THE GALACTIC BLACK HOLE

Lectures on general
relativity and astrophysics

EDITED BY
HEINO FALCKE
FRIEDRICH W HEHL

IoP

THE GALACTIC BLACK HOLE
Lectures on General Relativity and Astrophysics

Series in High Energy Physics, Cosmology and Gravitation

THE GALACTIC BLACK HOLE
Lectures on General Relativity and Astrophysics

Edited by

Heino Falcke

Max Planck Institute for Radio Astronomy,
Bonn, Germany

and

Friedrich W Hehl

Institute for Theoretical Physics,
University of Cologne, Germany

I*o*P

Contents

This page intentionally left blank

Preface

Evidence is accumulating that in the center of our own galaxy some 10^6 solar masses cluster in a region with a diameter of the order of a few astronomical units[1]. Theoretical analysis strongly suggests that this can only be a black hole. This is a gravitational configuration where the inner region is cut off from the outside by an event horizon, a semi-permeable closed surface surrounding it: material from the outside can fall in but communication from the inside to the outside is impossible.

Studies of other galaxies have shown that such supermassive black holes are rather common and probably reside at the center of every galaxy. Cosmologically speaking, the supermassive black hole in the Galactic Center is in our backyard, only about 26 000 light years away from us. This makes it the best observed candidate for studying all aspects of black hole physics and is an ideal laboratory for black hole physics.

The theory of black hole physics, developed mainly by general relativists and considered in the past as being no more than a mathematical curiosity, can now be applied to realistic astrophysical situations like that in our Galactic Center. Clearly, the time has come for general relativists and astrophysicists to collaborate on these issues and our book represents an attempt in this direction. The Galactic Center is a unique place where these two fields really start to touch each other.

On behalf of the German Physical Society (DPG) and jointly with Dr Joachim Debrus, director of the Physics Center in Bad Honnef, we organized a DPG School on the Galactic Black Hole in Bad Honnef addressing graduate students in physics, astronomy and mathematics from different countries. Whereas this was a school and not a workshop for specialists, we, nevertheless, invited as teachers physicists/astrophysicists who are working at the foremost research front of this subject. This book contains the lectures given at that school, in an order which should allow a beginner to tackle the material by commencing from fairly elementary topics in general relativity and in the astrophysics of our Galaxy right to the whereabouts of the central supermassive black hole. In fact, one of the goals, besides teaching the students, was to teach the scientists

---

[1] 1 AU $\simeq$ 150 $\times$ 10^6 km = average distance between earth and sun, see our table of units and constants in the back of the book, p 346.

themselves: astrophysics for the relativist and relativity for the astrophysicist. Hence, we hope the book will be a useful resource for students, lecturers and researchers in both fields alike.

<div align="right">

Heino Falcke (Bonn) and Friedrich W Hehl (Cologne)
August 2002

</div>

PART 1

GENERAL INTRODUCTION

This page intentionally left blank

Chapter 1

The Schwarzschild black hole: a general relativistic introduction

Christian Heinicke and Friedrich W Hehl
University of Cologne, Germany

The gravitational field of a homogeneous spherically symmetric body ('star') is derived in Newton's and in Einstein's gravitational theory, respectively. On the way to these results, Newton's theory is formulated in a quasi-field-theoretical form, its incompatibility with special relativity theory is pointed out, and it is outlined how one arrives at Einstein's field equation. The gravitational field of a simple Einsteinian model star consists of the exterior and the interior Schwarzschild solutions which are joined together at the surface of the star. Their derivation and interpretation will be discussed; in particular the Schwarzschild radius (for the sun ≈3 km) and its relation to the event horizon of the corresponding black hole will be investigated.

1.1 Newton's gravitational theory in quasi-field-theoretical form

> *Gravity exists in all bodies universally and is proportional to the quantity of matter in each ... If two globes gravitate towards each other, and their matter is homogeneous on all sides in regions that are equally distant from their centers, then the weight of either globe towards the other will be inversely as the square of the distance between the centers.*
>
> Isaac Newton (1687)

The gravitational force of a pointlike mass m_2 on a similar one of mass m_1 is given by Newton's attraction law:

$$F_{2\to 1} = -G\frac{m_1 m_2}{|r|^2}\frac{r}{|r|} \tag{1.1}$$

3
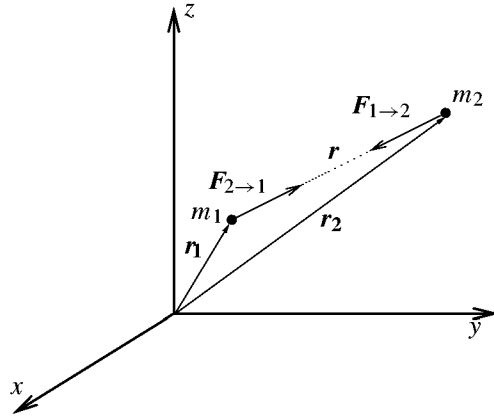
Figure 1.1. Two mass points m_1 and m_2 in three-dimensional space, Cartesian coordinates x, y, z.

where G is Newton's gravitational constant, see [8],

$$G \overset{\text{SI}}{=} 6.675\,59(27) \times 10^{-11} \frac{(\text{m/s})^4}{\text{N}}.$$

The vector $r := r_1 - r_2$ points from m_2 to m_1, see figure 1.1. According to *actio = reactio* (Newton's third law), we have $F_{2\to 1} = -F_{1\to 2}$. Thus complete symmetry exists in the gravitational interaction of the two masses with each other.

Let us now distinguish the mass m_2 as a field-generating active gravitational mass and m_1 as a (pointlike) passive test mass. Accordingly, we introduce a hypothetical *gravitational field* describing the force per unit mass ($m_2 \hookrightarrow M, m_1 \hookrightarrow m$):

$$f := \frac{F}{m} = -\frac{GM}{|r|^2} \frac{r}{|r|}. \tag{1.2}$$

With this definition, the force acting on the test mass m is equal to the *field strength* $\times$ *gravitational charge* (mass) or $F_{M\to m} = mf$, in analogy to electrodynamics. The active gravitational mass M is thought to emanate a gravitational field which is always directed to the center of M and has the same magnitude on every sphere with M as center, see figure 1.2.

Let us now investigate the properties of the gravitational field (1.2). Obviously, there exists a potential

$$\phi = -G\frac{M}{|r|} \qquad f = -\operatorname{grad}\phi. \tag{1.3}$$

Accordingly, the gravitational field is curl free: $\operatorname{curl} f = 0$.
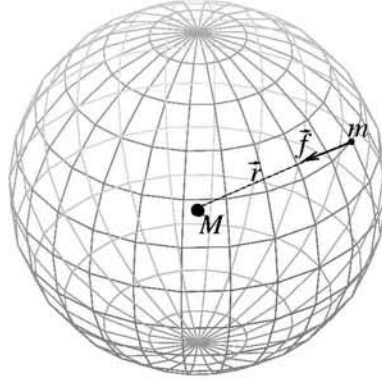
Figure 1.2. The 'source' M attracts the test mass m.

By assumption it is clear that the source of the gravitational field is the mass M. We find, indeed, that

$$\text{div } f = -4\pi G M \delta^3(r) \tag{1.4}$$

where $\delta^3(r)$ is the three-dimensional (3D) delta-function. By means of the *Laplace operator* $\Delta := \text{div grad}$, we infer, for the gravitational potential, that

$$\Delta\phi = 4\pi G M \delta^3(r). \tag{1.5}$$

The term $M\delta^3(r)$ may be viewed as the mass density of a point mass. Equation (1.5) is a second-order linear partial differential equation for ϕ. Thus the gravitational potential generated by several point masses is simply the linear superposition of the respective single potentials. Hence we can generalize the *Poisson equation* (1.5) straightforwardly to a continuous matter distribution $\rho(r)$:

$$\Delta\phi = 4\pi G\rho. \tag{1.6}$$

This equation interrelates the source ρ of the gravitational field with the gravitational potential ϕ and thus completes the quasi-field-theoretical description of Newton's gravitational theory.

We speak here of *quasi*-field-theoretical because the field ϕ as such represents a convenient concept. However, it has no *dynamical* properties, no genuine degrees of freedom. The Newtonian gravitational theory is an *action at a distance* theory. When we remove the source, the field vanishes instantaneously. Newton himself was very unhappy about this consequence. Therefore he emphasized the preliminary and purely descriptive character of his theory. But before we liberate the gravitational field from this constraint by equipping it with its own degrees of freedom within the framework of general relativity theory, we turn to some properties of the Newtonian theory.
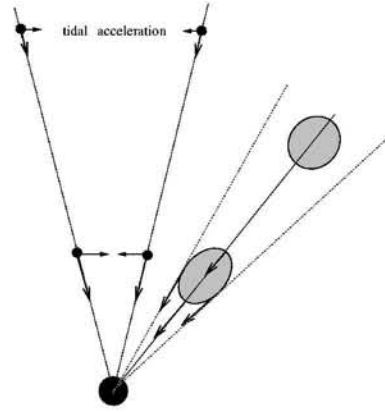
Figure 1.3. Tidal forces in a spherically symmetric gravitational field.

A very peculiar fact characteristic to the gravitational field is that the acceleration of a freely falling test body does not depend on the mass of this body but only on its position within the gravitational field. This comes about because of the equality (in suitable units) of the gravitational and inertial mass:

$$\overset{\text{inertial}}{m}\,\ddot{r} = F = \overset{\text{grav}}{m}\,f. \tag{1.7}$$

This equality has been well tested since Galileo's time by means of pendulum and other experiments with an ever increasing accuracy, see Will [21].

In order to allow for a more detailed description of the structure of a gravitational field, we introduce the concept of a *tidal force*. This can be best illustrated by means of figure 1.3. In a spherically symmetric gravitational field, for example, two test masses will fall radially towards the center and thereby get closer and closer. Similarly, a spherical drop of water is deformed to an ellipsoidal shape because the gravitational force at its bottom is bigger than at its top, which is at a greater distance from the source. If the distance between two freely falling test masses is relatively small, we can derive an explicit expression for their relative acceleration by means of a Taylor expansion. Consider two mass points with position vectors r and $r + \delta r$, with $|\delta r| \ll 1$. Then the relative acceleration reads:

$$\Delta a = [f(r + \delta r) - f(r)] = \delta r \cdot \text{Grad}\, f(r) \tag{1.8}$$

where Grad denotes the vector gradient. We may rewrite this accordingly (the sign is conventional, $\partial/\partial x^a =: \partial_\alpha$, $x^1 = x$, $x^2 = y$, $x^3 = z$):

$$K_{ab} := -(\text{Grad}\, f)_{ab} = -\partial_a f_b \qquad a, b = 1, 2, 3.$$

We call K_{ab} the *tidal force* matrix. The vanishing curl of the gravitational field is equivalent to the symmetry $K_{ab} = K_{ba}$. Furthermore, $K_{ab} = \partial_a \partial_b \phi$. Thus, the

Poisson equation becomes

$$\sum_{a=1}^{3} K_{aa} = \text{trace } K = 4\pi G\rho. \tag{1.9}$$

Accordingly, in vacuum K_{ab} is trace free.

Let us now investigate the gravitational potential of a homogeneous star with constant mass density $\rho_\odot$ and total mass $M_\odot = (4/3)\pi R_\odot^3 \rho_\odot$. For our sun, the radius is $R_\odot = 6.9598 \times 10^8$ m and the total mass is $M = 1.989 \times 10^{30}$ kg.

Outside the sun (in the idealized picture we are using here), we have a vacuum. Accordingly, $\rho(\mathbf{r}) = 0$ for $|\mathbf{r}| > R_\odot$. Then the Poisson equation reduces to the *Laplace equation*

$$\Delta\phi = 0 \qquad \text{for } r > R_\odot. \tag{1.10}$$

In 3D polar coordinates, the r-dependent part of the Laplacian has the form $(1/r^2)\partial_r(r^2\partial_r)$. Thus (1.10) has the solution

$$\phi = \frac{\alpha}{r} + \beta \tag{1.11}$$

where α and β are integration constants. Requiring that the potential tends to zero as r goes to infinity, we get $\beta = 0$. The integration constant α will be determined from the requirement that the force should change smoothly as we cross the star's surface, i.e. the interior and exterior potentials and their first derivatives have to be matched continuously at $r = R_\odot$.

Inside the star we have to solve

$$\Delta\phi = 4\pi G\rho_\odot \qquad \text{for } r \leq R_\odot. \tag{1.12}$$

We find

$$\phi = \frac{2}{3}\pi G\rho_\odot r^2 + \frac{C_1}{r} + C_2 \tag{1.13}$$

with integration constants C_1 and C_2. We demand that the potential in the center $r = 0$ has a finite value, say ϕ_0. This requires $C_1 = 0$. Thus

$$\phi = \frac{2}{3}\pi G\rho_\odot r^2 + \phi_0 = \frac{GM(r)}{2r} + \phi_0 \tag{1.14}$$

where we have introduced the *mass function* $M(r) = (4/3)\pi r^3 \rho_\odot$ which measures the total mass inside a sphere of radius r.

Continuous matching of ϕ and its first derivatives at $r = R_\odot$ finally yields:

$$\phi(\mathbf{r}) = \begin{cases} -G\dfrac{M_\odot}{|\mathbf{r}|} & \text{for } |\mathbf{r}| \geq R_\odot \\[2mm] G\dfrac{M_\odot}{2R_\odot^3}|\mathbf{r}|^2 - \dfrac{3GM_\odot}{2R_\odot} & \text{for } |\mathbf{r}| < R_\odot. \end{cases} \tag{1.15}$$

The slope (first derivative) and the curvature (second derivative) of this curve represent the magnitudes of the gravitational and the tidal forces, respectively.
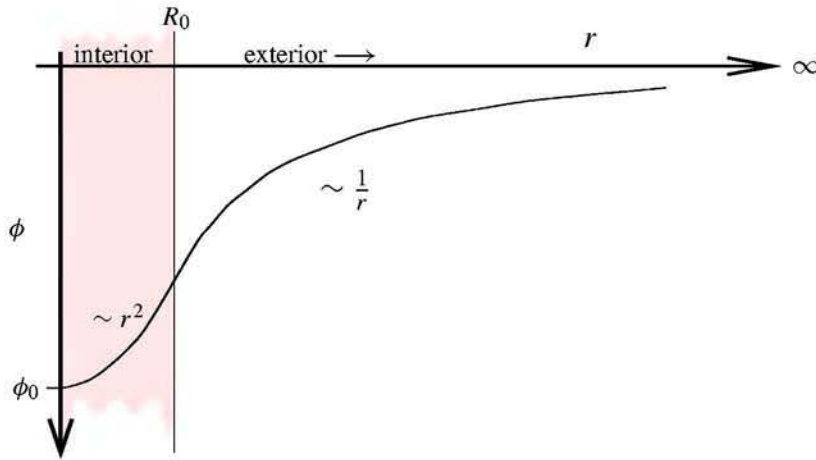
Figure 1.4. Newtonian potential of a homogeneous star.

1.2 Special relativity and Newton's theory: a clash

Not only have we no direct intuition of the equality of two periods, but we have not even direct intuition of the simultaneity of two events oc-curring in two different places.

Henri Poincaré (1902)

Apparently, the space surrounding us has three dimensions. Together with the one-dimensional time, it constitutes *four-dimensional* (4D) *spacetime*. Distinguished frames of reference are the inertial frames. They are understood as infinitely extended frames in which *force-free* particles are at rest or move uniformly along straight lines in the sense of Euclidean geometry. In them, we may introduce coordinates

$$x^0 = ct, x^1 = x, x^2 = y, x^3 = z, \text{ or } x^\mu \qquad \text{with } \mu = 0, 1, 2, 3. \qquad (1.16)$$

As a rule, all Greek indices shall run from 0 to 3. In an empty space with respect to an *inertial* frame of reference, there is no distinction between the different points in it and no preferred direction. Likewise, there is no preferred instant of time.

With this homogeneous and spatially isotropic spacetime in mind, we state the *special relativity principle:* the laws of physics are the same in all inertial frames.

A prototypical law of nature to be stated in this context is the *principle of the constancy of the speed of light:* light signals in vacuum are propagated

rectilinearly, with the same speed c at all times, in all directions, in all inertial frames, independently of the motion of their sources.

By means of these two principles, we can deduce the *Poincaré* (or inhomogeneous Lorentz) transformations which encompass four spacetime translations, three spatial rotations, and three *Lorentz boosts*, i.e. velocity transformations. The 'essence' of this transformation can also be expressed in a somewhat different manner.

We define a tensor **T** of covariant rank k and contravariant rank l, respectively, by means of its behavior under coordinate transformations,

$$T^{\mu_1' \dots \mu_l'}{}_{\nu_1' \dots \nu_k'} = P^{\mu_1'}_{\mu_1} \dots P^{\mu_l'}_{\mu_l} P^{\nu_1}_{\nu_1'} \dots P^{\nu_k}_{\nu_k'} T^{\mu_1 \dots \mu_l}{}_{\nu_1 \dots \nu_k} \tag{1.17}$$

where we have introduced the *Jacobian* matrix and its inverse according to

$$P^{\alpha'}_{\alpha} := \frac{\partial x^{\alpha'}}{\partial x^{\alpha}} \qquad P^{\alpha}_{\alpha'} = \frac{\partial x^{\alpha}}{\partial x^{\alpha'}} \qquad P^{\alpha}_{\alpha'} P^{\alpha'}_{\beta} = \delta^{\alpha}_{\beta}. \tag{1.18}$$

The summation convention is assumed, i.e. summation is understood over repeated indices. The values of the components of tensors do change, but only in the specific linear and homogeneous manner indicated here. Equations of tensors remain form invariant or covariant, i.e. the transformed equations look the same but with the unprimed indices replaced by primed ones. If one *contracts* co- and contravariant tensors in such a way that no free index is left, $v_i w^i$, e.g. one gets a scalar, which is *invariant* under transformations, i.e. it does not change its value. The latter represents an observable quantity. The generic case of a covariant tensor of first rank is the partial derivative of a scalar function $\phi_{,\alpha} := \partial \phi / \partial x^{\alpha}$ and the typical contravariant tensor is the coordinate differential dx^{α}. Besides tensors, we also need spinors in special relativity, but they are not essential in gravitational theory.

We define the *Minkowski metric* as follows.

$$ds^2 := -c^2 \, dt^2 + dx^2 + dy^2 + dz^2 = g_{\alpha\beta} \, dx^{\alpha} \, dx^{\beta} \tag{1.19}$$

where (in Cartesian coordinates)

$$g_{\alpha\beta} \overset{*}{=} \eta_{\alpha\beta} := \text{diag}(-1, +1, +1, +1) = \eta^{\alpha\beta} \overset{*}{=} g^{\alpha\beta}. \tag{1.20}$$

The $g^{\alpha\beta}$ denote the inverse of the metric tensor. Under a Poincaré transformation, the components of the Minkowski metric $\eta_{\alpha\beta}$ remain numerically invariant. This metric defines an invariant spatiotemporal distance between two spacetime points or *events*, as they are called. Spatial distance alone between two points can be different for different observers and the same applies to time intervals. This manifests itself in the well-known effects of time dilation and length contraction.

Now we are able to express the principle of special relativity in the following way: the equations of physics describing laws of nature transform covariantly under Poincaré transformations.

How can we apply this to gravity? In Newtonian gravity, the potential obeys the Poisson equation $\Delta\phi = 4\pi G\rho$. The corresponding wave equation can be represented as

$$\Box\phi = \partial_\alpha(\eta^{\alpha\beta}\partial_\beta\phi) = -\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} + \Delta\phi = 4\pi G\rho \qquad (1.21)$$

and thus is manifestly Poincaré invariant. Hence, the Poisson equation as such is *not* Poincaré invariant but only a limiting case of the wave equation for static situations.

The first idea for a Poincaré-covariant equation for the gravitational potential would be the obvious generalization by admitting the gravitational potential ϕ and the source ρ to be time dependent and interrelating both by means of a gravitational wave equation $\Box\phi = 4\pi G\rho$. But what is the source ρ now? In the case of a pressure-less fluid or a swarm of dust particles where all components move parallelly with the same velocity (and correspondingly have a common rest system), a Poincaré-invariant meaning for the mass density can be found, but this is not possible in general. Moreover, we learn from special relativity that mass and energy are equivalent. Binding forces and therewith stress within matter are expected to contribute to its gravitating mass. Thus, in a relativistic theory of gravitation, we have to replace mass density by energy density. Next, we have to look for a Poincaré invariant quantity which contains the (mass-)energy density and will reduce to it in special cases.

And indeed, special relativity provides such a quantity. In electrodynamics, Minkowski found a symmetric second-rank tensor $T^{\alpha\beta}_{\text{Max}}$ whose divergence yields the Lorentz force density $\partial_\alpha T^{\alpha\beta}_{\text{Max}} = f^\beta$. For an electrically charged perfect fluid, characterized by mass-energy density ρ and pressure p, the equations of motion can be written in the form

$$\partial_\alpha(T^{\alpha\beta}_{\text{Max}} + T^{\alpha\beta}_{\text{Mat}}) = 0 \qquad (1.22)$$

where we have introduced the *energy–momentum tensor* of the perfect fluid:

$$T^{\alpha\beta}_{\text{Mat}} = \left(\rho + \frac{p}{c^2}\right)u^\alpha u^\beta + pg^{\alpha\beta}. \qquad (1.23)$$

The vector $u^\alpha = dx^\alpha/d\tau = \gamma(v)(c, \mathbf{v})$ is the four-velocity of the fluid elements (and $\mathbf{v}$ the three-velocity with respect to the chosen frame of reference. The *Lorentz factor* γ is given by $\gamma(v) := (1 - v^2/c^2)^{-1/2}$). The components of the energy–momentum tensor are not invariant, of course. In the rest frame of the fluid, the observer sees a fluid at rest with a certain mass-energy distribution and an isotropic pressure p: $T^{\alpha\beta} \overset{*}{=} \text{diag}(\rho c^2, p, p, p)$. However, with respect to a moving frame, there is a moving energy distribution which results in an energy flux density. Moreover, isotropic pressure transforms into anisotropic stress etc. In general, we arrive at the following structure (momentum flux density and stress

are equivalent notions, $i, j = 1, 2, 3$):

$$T_{\mu\nu} = \left(\begin{array}{c|c} T_{00} & T_{0i} \\ \hline T_{i0} & T_{ij} \end{array} \right) = \left(\begin{array}{c|c} \begin{array}{c} \text{energy} \\ \text{density} \end{array} & \text{momentum density} \\ \hline \begin{array}{c} \text{energy} \\ \text{flux} \\ \text{density} \end{array} & \begin{array}{c} \text{momentum} \\ \text{flux} \\ \text{density} \end{array} \end{array} \right). \tag{1.24}$$

Now we can construct a scalar invariant encompassing the mass-energy density in the following way:

$$T := T_\alpha{}^\alpha = g_{\alpha\beta} T^{\alpha\beta} = -\rho c^2 + 3p. \tag{1.25}$$

For 'non-relativistic matter', we find $\rho \ll 3p/c^2$. Thus, indeed, $T \approx \rho c^2$. The Poincaré-invariant field equation

$$\Box \phi = \kappa T \tag{1.26}$$

then yields the Newtonian Poisson equation in an appropriate limiting case and for an appropriately chosen coupling constant κ.

At first sight, this defines a viable gravitational theory. However, it turns out that this theory runs into serious conflicts with observations. A scalar gravitational theory does not allow for the deflection of light in gravitational fields because a scalar field cannot be coupled reasonably to the electromagnetic field, since the electromagnetic energy–momentum tensor is traceless. Light deflection has been experimentally confirmed beyond doubt. Thus, we have to look for different possibilities in order to interrelate electromagnetic energy–momentum and the gravitational potential. To this end we will now turn to the gravitational field.

1.3 Accelerated frames of reference, equivalence principle and Einstein's field equation

Die Relativitätstheorie bringt uns aber nicht nur den Zwang, Newtons Theorie zu modifizieren; sie schränkt auch zum Glück in weitgehendem Maße die Möglichkeiten hierfür ein.

Albert Einstein (1913)

An observer who measures the acceleration of a freely falling body within a sufficiently small laboratory obtains the same results whether his/her laboratory is at rest in a gravitational field or appropriately accelerated in gravity-free space. Consequently, the quantity representing the inertial forces in the equation of motion should be similar to the quantity representing the gravitational forces. In an inertial frame in Cartesian coordinates x^μ, a force-free test particle obeys the equation of motion

$$m \frac{d^2 x^\mu}{d\tau^2} = 0. \tag{1.27}$$
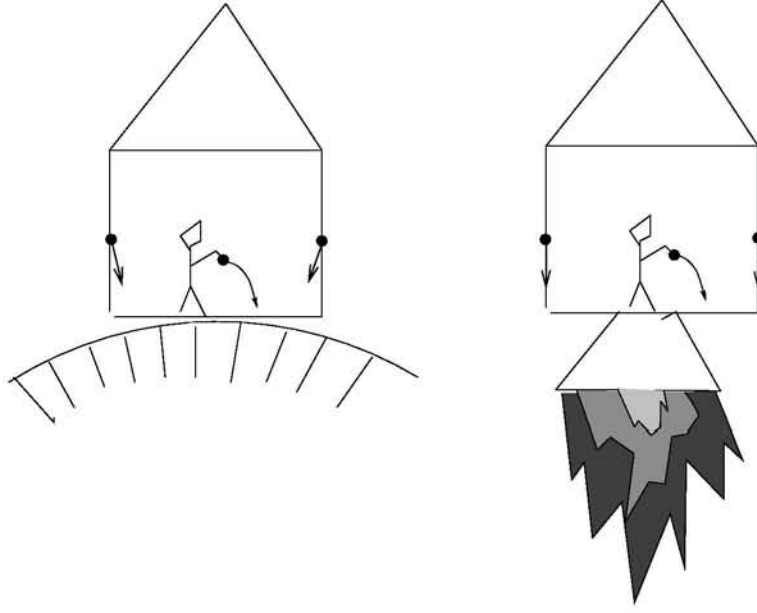
Figure 1.5. The local equivalence of an accelerated frame of reference and a gravitational field. Note, if we compare the gravitational and the inertial forces acting on *two* point particles in each case, because of the tidal effect, we can distinguish the laboratory on earth and that in space. However, locally, *one* test particle moves in the same way in both laboratories.

Thus it moves in a straight line $x^\mu(\tau) = a^\mu + b^\mu \tau$ (a^μ, b^μ constant vectors). The space laboratory represents an accelerated frame of reference with coordinates $x^{\mu'}$. We apply a coordinate transformation $x^{\alpha'}(x^\mu)$ to (1.27) and find

$$m\frac{d^2 x^{\alpha'}}{d\tau^2} + m\Gamma^{\alpha'}{}_{\beta'\gamma'}\frac{dx^{\beta'}}{d\tau}\frac{dx^{\gamma'}}{d\tau} = 0 \qquad (1.28)$$

where the *connection* components

$$\Gamma^{\alpha'}{}_{\beta'\gamma'} \stackrel{*}{=} \frac{\partial x^{\alpha'}}{\partial x^\alpha}\frac{\partial^2 x^\alpha}{\partial x^{\beta'}\partial x^{\gamma'}} \qquad (1.29)$$

represent the *inertial field*. For a rotating coordinate system, e.g. Γ encompasses the Coriolis force etc. So far $\Gamma^{\alpha'}{}_{\beta'\gamma'}$ is only an coordinate artifact and has no degrees of freedom of its own. We can always introduce a global coordinate system such that the $\Gamma^{\alpha'}{}_{\beta'\gamma'}$ vanish everywhere.

We can deduce an alternative representation of $\Gamma^{\alpha'}{}_{\beta'\gamma'}$ from the tensorial transformation behavior of the metric tensor (we suppress the dashes here):

$$\Gamma^\alpha{}_{\mu\nu} := \tfrac{1}{2}g^{\alpha\beta}(\partial_\nu g_{\beta\mu} + \partial_\mu g_{\beta\nu} - \partial_\beta g_{\mu\nu}). \qquad (1.30)$$

Thus, the connection components, also called the *Christoffel symbols* in the case of a Riemannian space, can be expressed in terms of ten functions $g_{\alpha\beta} = g_{\beta\alpha}$ which tentatively serve as the gravitational or inertial potential. In order to be able to choose a coordinate system such that $\Gamma^{\alpha'}{}_{\beta'\gamma'} = 0$ globally, the $\Gamma^{\alpha}{}_{\beta\gamma}$ have to fulfil a certain integrability condition, namely their 'curl' has to vanish:

$$0 = R^{\mu}{}_{\nu\alpha\beta} := \partial_{\alpha}\Gamma^{\mu}{}_{\nu\beta} - \partial_{\beta}\Gamma^{\mu}{}_{\nu\alpha} + \Gamma^{\mu}{}_{\sigma\alpha}\Gamma^{\sigma}{}_{\nu\beta} - \Gamma^{\mu}{}_{\sigma\beta}\Gamma^{\sigma}{}_{\nu\alpha}. \qquad (1.31)$$

The quantity $R^{\alpha}{}_{\beta\mu\nu}$ is called the *Riemannian curvature* tensor. If $R^{\alpha}{}_{\beta\mu\nu} = 0$, we have a flat Minkowski spacetime (possibly in curvilinear coordinates), whereas $R^{\alpha}{}_{\beta\mu\nu} \neq 0$ implies a *curved Riemannian spacetime*. In a Riemannian space, the curvature tensor fulfills certain *algebraic* identities which reduce its number of independent components to 20:

$$R_{\alpha\beta\mu\nu} = -R_{\alpha\beta\nu\mu}, \quad R_{\alpha\beta\mu\nu} = -R_{\beta\alpha\mu\nu}, \quad R_{\alpha\beta\mu\nu} + R_{\alpha\nu\beta\mu} + R_{\alpha\mu\nu\beta} = 0. \qquad (1.32)$$

Let us now construct the field equation for gravity by trying to proceed along the same line as in other successful field theories, such as electrodynamics. The equations of motion with the abbreviation $(\dot{\ }) = d/d\tau$ read:

$$
\begin{array}{cccccc}
\text{Maxwell:} & m\,\ddot{x}^{\alpha} & = & q\,\dot{x}^{\mu} & \times & F^{\alpha}{}_{\mu} \\
& & & \text{electric} & & \text{el.-mag.} \\
& & & \textit{current} & & \textit{field strength} \\
& & & \text{inertial} & & \text{inertial} \\
\text{Gravitation:} & m\,\ddot{x}^{\alpha} & = & -m\,\dot{x}^{\mu}\,\dot{x}^{\nu} & \times & \Gamma^{\alpha}{}_{\mu\nu}.
\end{array}
\qquad (1.33)
$$

This fits quite nicely into our considerations in the previous section. The current, which couples to the inertial field, is the quantity $m\dot{x}^{\mu}\dot{x}^{\nu}$ which corresponds to the energy–momentum tensor of dust $T^{\alpha\beta} = \rho\dot{x}^{\alpha}\dot{x}^{\beta}$. This coincides with the earlier suggestion that $T^{\alpha\beta}$ should be the source of gravity.

In electrodynamics, we have the four-potential $A_{\mu} = (\phi_{\text{elec}}, \boldsymbol{A})$, ϕ_{elec} is the 3D scalar electric potential, $\boldsymbol{A}$ the 3D vector potential. Furthermore, the electromagnetic field strength is denoted by $F_{\alpha\beta} = \partial_{\alpha}A_{\beta} - \partial_{\beta}A_{\alpha}$ and the current by J^{α}. With the *Lorenz gauge*, $\partial_{\mu}A^{\mu} = 0$, we find that

$$
\begin{array}{ccccc}
\text{divergence of } \textit{field} & \sim & \text{d'Alembertian on } \textit{potential} & \sim & \textit{source current} \\
\partial_{\mu}F^{\mu\nu} & = & \Box A^{\nu} & = & J^{\nu}.
\end{array}
\qquad (1.34)
$$

However, it is not so simple in gravity. Gravitational radiation carries energy, and energy is, as we have argued earlier, itself a source of gravity. Thus, the gravitational field has a self-interaction which distinguishes it from the electromagnetic field. Consequently, gravity is described by a nonlinear field equation of the following type:

$$\text{'Div'}\,\Gamma + \Gamma^2 \sim \Box g_{\alpha\beta} + \text{nonlinear} \sim T_{\alpha\beta}. \qquad (1.35)$$

That the nonlinearity is only quadratic will be a result of our subsequent considerations.

So much for the general outline. To fix an exact equation, we need some additional criteria. In particular, we have to say something about general covariance. We consider an accelerated frame of reference locally equivalent to one which is at rest in a gravitational field. Gravity is a relatively weak force, but it has an infinite range and is all pervading. We will hardly find a gravity-free spot in the universe. Hence, in general we find ourselves in a non-inertial frame, even if the deviation from an inertial system may be negligible on small scales. From this point of view, the fundamental laws of physics should be covariant not only under Poincaré transformations but also under general coordinate transformations. There is not much change with respect to the algebra of tensors, but a very noticeable change comes about in tensor analysis: the partial derivative of a tensor will not transform like a tensor. This can be fixed by introducing the so-called *covariant derivative*:

$$\nabla_\alpha T^\mu{}_\nu = \partial_\alpha T^\mu{}_\nu + \Gamma^\mu{}_{\gamma\alpha} T^\gamma{}_\nu - \Gamma^\gamma{}_{\nu\alpha} T^\mu{}_\gamma. \tag{1.36}$$

By replacing the partial derivatives in the special relativistic formulae by covariant ones, we obtain general covariant equations. This 'correspondence' principle mostly, but not always, yields physical reasonable generalizations of the special relativistic laws.

In Newton's theory, the mass density as source is linearly related to the tidal force. Can we also define tidal forces in general relativity?

The equation of motion (1.27) has a geometrical interpretation, too. The metric allows the definition of an invariant length of a curve γ, parametrized by $x^\mu = x^\mu(\tau)$, connecting two spacetime points $\mathsf{A} = x^\mu(0)$ and $\mathsf{B} = x^\mu(\tau_0)$ by means of the line integral

$$l = \int_\gamma ds = \int_0^{\tau_0} d\tau \sqrt{\dot{x}^\mu \dot{x}^\nu g_{\mu\nu}}. \tag{1.37}$$

This length l represents the proper time of an observer who moves along the path γ from A to B. The necessary and sufficient condition for γ to be a curve of extremal length is found to be (provided γ is parametrized by its arc length)

$$\ddot{x}^\mu + \Gamma^\mu{}_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta = 0. \tag{1.38}$$

This is the *Euler–Lagrange* equation for the variational problem $\delta \int ds = 0$; and it coincides with the equation of motion (1.27). In geometry, (1.38) is called the *geodesic equation* and its solutions x^μ are *geodesics*. In flat space, the geodesics are straight lines, the geodesics of a sphere are circuits, etc.

Thus, freely falling particles move along the geodesics of Riemannian spacetime. Now we can address the question of tidal accelerations between two freely falling particles. Let the vector v^μ be the vector describing the distance

between two particles moving on infinitesimally adjacent geodesics. A simple calculation yields the equation for the *geodesic deviation*:

$$\frac{D^2 v^\mu}{D\tau^2} = \dot{x}^\nu \dot{x}^\alpha v^\beta R^\mu{}_{\nu\alpha\beta} \tag{1.39}$$

where $D/D\tau$ denotes the absolute derivative along the curve x^α. Eventually, the tidal acceleration is represented by the curvature tensor. In Newton's theory, the tidal force is linearly related to the tidal acceleration. The energy–momentum tensor, as the suspected source of gravity, is a symmetric second-rank tensor. Therefore it has ten independent components.

Now only the problem of how to interrelate the second-rank symmetric energy–momentum tensor to the fourth-rank Riemannian curvature tensor remains. In analogy to the Newtonian case we would like this relation to be linear. It turns out that such a relation has to be of the form

$$\alpha R_{\mu\nu} + \beta R g_{\mu\nu} = T_{\mu\nu} \tag{1.40}$$

with the *Ricci tensor*

$$R_{\alpha\beta} := R^\mu{}_{\alpha\mu\beta} \tag{1.41}$$

and the *curvature scalar*

$$R := R^\alpha{}_\alpha. \tag{1.42}$$

The constants α and β have to be fixed by additional conditions. The vanishing divergence of the energy–momentum together with the second Bianchi identity (a kind of integrability condition)

$$\nabla_\lambda R^\alpha{}_{\beta\mu\nu} + \nabla_\nu R^\alpha{}_{\beta\lambda\mu} + \nabla_\mu R^\alpha{}_{\beta\nu\lambda} = 0 \tag{1.43}$$

leads to *Einstein's field equation*:

$$\underbrace{R_{\mu\nu} - \tfrac{1}{2} R g_{\mu\nu}}_{\text{Einstein tensor } G_{\mu\nu}} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}. \tag{1.44}$$

The value $\kappa := 8\pi G/c^4$ of *Einstein's gravitational constant* can be determined by a transition to the Newtonian limit of general relativity. Moreover, we have added a cosmological term containing the *cosmological constant* Λ.

The energy–momentum tensor has ten independent components whereas the Riemannian curvature tensor has 20 independent components. Hence, the energy–momentum tensor determines only a part of the curvature. Indeed, we have the decomposition

$$R_{\mu\nu\alpha\beta} = C_{\mu\nu\alpha\beta} + \tfrac{1}{2}(g_{\mu\alpha} L_{\beta\nu} - g_{\mu\beta} L_{\alpha\nu} - g_{\nu\alpha} L_{\beta\mu} + g_{\nu\beta} L_{\alpha\mu}) \tag{1.45}$$

where

$$L_{\alpha\beta} := R_{\alpha\beta} - \tfrac{1}{6} R g_{\alpha\beta} = L_{\beta\alpha} \tag{1.46}$$

(for recent work on the L-tensor, see [11]). This part of the curvature is algebraically linked to the matter distribution by means of Einstein's equation. Consequently, it vanishes in vacuum and only the irreducible fourth-rank part $C_{\alpha\beta\gamma\delta}$, the conformal *Weyl curvature* with ten independent components, remains.

1.4   The exterior Schwarzschild solution

Es ist eine ganz wunderbare Sache, dass von einer so abstrakten Idee
aus die Erklärung der Merkuranomalie so zwingend herauskommt.

Karl Schwarzschild (1915)

Just a few months after Einstein had published his new gravitational theory, the astronomer K Schwarzschild found an exact solution to Einstein's field equation. The so-called Schwarzschild solution is amazingly simple, especially in view of the field equations which are very complicated. However, the Schwarzschild solution is not a degenerated case for over-simplified situations but physically most meaningful. It is this solution by means of which one can explain most general relativistic effects in the planetary system. The reason is that it describes the gravitational field outside of a spherically symmetric body— like the planets and the sun.

We start from the spherically symmetric metric:

$$ds^2 = -e^{\nu(r,t)}c^2\,dt^2 + e^{\lambda(r,t)}\,dr^2 + r^2\,d\Omega^2 \qquad d\Omega^2 := d\theta^2 + \sin^2\theta\,d\varphi^2. \quad (1.47)$$

One can now compute the Christoffel symbols, the Riemann tensor, and the Einstein tensor for this ansatz. This can be done by hand, of course. It is more convenient to use computer algebra, see section 1.8. For vacuum and $\Lambda = 0$, it is relatively simple to find a solution to $G_{\alpha\beta} = \kappa T_{\alpha\beta} = 0$, namely

$$ds^2 = -\left(1 - \frac{2m}{r}\right)c^2\,dt^2 + \frac{1}{1 - \frac{2m}{r}}\,dr^2 + r^2\,d\Omega^2. \quad (1.48)$$

This is the *Schwarzschild metric* [19]. There is no time dependence although we did allow for that in the ansatz (1.47). The vacuum spacetime structure generated by any spherically symmetric body is *static*. This applies also for the exterior field of a radially oscillating body. This fact is known as *Birkhoff's theorem*.

The parameter $2m$ is an integration constant. Its interpretation can be obtained by means of a transition to Newton's theory. It turns out that (G is Newton's gravitational constant and M is the mass of gravitating body)

$$r_S := 2m = \frac{2GM}{c^2}. \quad (1.49)$$

This is the *Schwarzschild* (or gravitational) *radius*. In this chapter, we distinguish between m and M. In subsequent chapters, it is generally assumed that $c = 1 =$

G. Then, we have $m = M$. At the *Schwarzschild radius* r_S the metric coefficients become singular. However, this is only a so-called *coordinate singularity* since the curvature tensor (and therewith physically meaningful quantities like the tidal force) remains finite. We can also see this explicitly when we introduce suitable coordinates, like *isotropic coordinates*. Therefore we define a new radial coordinate $\bar{r}$ as follows:

$$r = \bar{r}\left(1 + \frac{m}{2\bar{r}}\right)^2.$$

(1.50)

Then, the Schwarzschild metric becomes

$$ds^2 = \left(\frac{1 - \frac{m}{2\bar{r}}}{1 + \frac{m}{2\bar{r}}}\right)^2 c^2\,dt^2 - \left(1 + \frac{m}{2\bar{r}}\right)^4 (d\bar{r}^2 + \bar{r}^2\,d\Omega^2).$$

(1.51)

In these coordinates, there is only a singularity at $\bar{r} = 0$, which corresponds to $r = 0$.

As already indicated at the beginning of this section, several experimental verifications of general relativity theory rest on the exterior Schwarzschild solution, namely, to mention only some of the catchwords,

- the gravitational red shift,
- the gravitational deflection of light ($\rightarrow$ gravitational lensing),
- the general relativistic perihelion and periastron advance, and
- the time delay of radar pulses (the Shapiro effect).

Using additional structure from Einstein's theory, more predictions can be verified:

- the Hulse–Taylor pulsar: emission of gravitational waves,
- the Lense–Thirring effect (see Ciufolini *et al* [2, 3] and Everitt [6]).

For more details on the experimental verification of Einstein's theory, see Will [21].

1.5 Flat Minkowski spacetime, null coordinates, and the Penrose diagram

In this section, we are going to analyze the Schwarzschild solution, in particular its singularity structure. For this purpose we will first have a look at null coordinates. The simplest testing ground in this context is the (flat) Minkowski space. Its metric, in Cartesian and spherical polar coordinates, reads ($c = 1$) as

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 = -dt^2 + dr^2 + r^2\,d\Omega^2.$$

(1.52)

We define *advanced* and *retarded null coordinates* as follows

$$v := t + r \qquad u := t - r$$

(1.53)
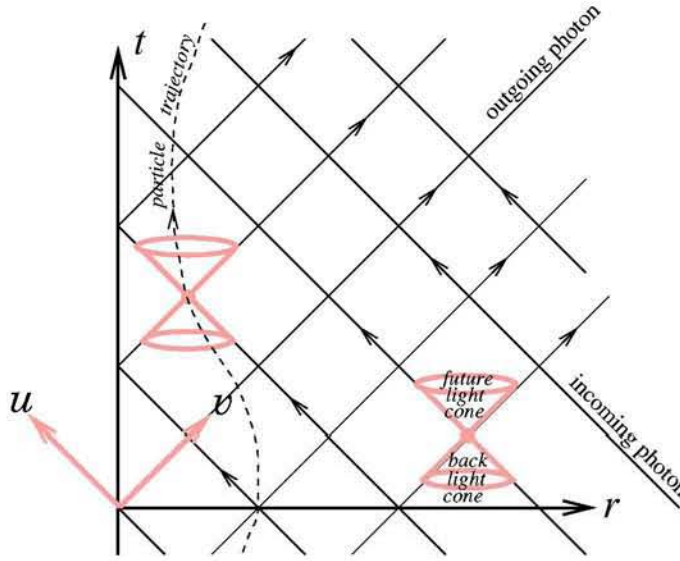
Figure 1.6. Minkowski spacetime in null coordinates.

and find that

$$ds^2 = -dv\,du + \tfrac{1}{4}(v - u)^2\,d\Omega^2. \tag{1.54}$$

In figure 1.6 we show the Minkowski spacetime in terms of the new coordinates. Incoming photons, i.e. pointlike particles with velocity $\dot{r} = -c = -1$, move on paths with $v = $ constant. Correspondingly, we have for outgoing photons $u = $ constant. The special relativistic wave equation is solved by any function $f(u)$ and $f(v)$. The surfaces $f(u) = $ constant and $f(v) = $ constant represent the wavefronts which evolve with the velocity of light. The trajectory of every material particle with $v < c = 1$ has to remain inside the region defined by the surface $r = t$. In an (r, t)-diagram this surface is represented by a cone, the so-called *light cone*. Any point in the *future light cone* $r = t$ can be reached by a particle or signal with a velocity less than c. A given spacetime point P can be reached by a particle or signal from the spacetime region enclosed by the *back light cone* $r = -t$.

We can map, following Penrose, the infinitely distant points of spacetime into finite regions by means of a conformal transformation which leaves the light cones intact. Then we can display the whole infinite Minkowski spacetime on a (finite) piece of paper. Accordingly, introduce the new coordinates

$$\tilde{v} := \arctan v \qquad \tilde{u} := \arctan u \qquad \text{for } -\pi/2 \le \tilde{v}, \tilde{u} \le +\pi/2. \tag{1.55}$$

Then the metric reads as

$$ds^2 = \frac{1}{\cos^2 \tilde{v}} \frac{1}{\cos^2 \tilde{u}} \left[-d\tilde{v}\,d\tilde{u} + \frac{1}{4} \sin^2(\tilde{v} - \tilde{u})\,d\Omega^2 \right]. \qquad (1.56)$$

We can go back to time- and spacelike coordinates by means of the transformation

$$\tilde{t} := \tilde{v} + \tilde{u} \qquad \tilde{r} := \tilde{v} - \tilde{u} \qquad (1.57)$$

see (1.52). Then the metric reads as

$$ds^2 = \frac{-d\tilde{t}^2 + d\tilde{r}^2 + \sin^2 \tilde{r}\,d\Omega^2}{4\cos^2 \frac{\tilde{t}+\tilde{r}}{2} \cos^2 \frac{\tilde{t}-\tilde{r}}{2}} \qquad (1.58)$$

that is, up to the function in the denominator, it appears as a flat metric. Such a metric is called conformally flat (it is conformal to a static Einstein cosmos). The back-transformation to our good old Minkowski coordinates reads:

$$t = \frac{1}{2} \left(\tan \frac{\tilde{t}+\tilde{r}}{2} + \tan \frac{\tilde{t}-\tilde{r}}{2} \right) \qquad (1.59)$$

$$r = \frac{1}{2} \left(\tan \frac{\tilde{t}+\tilde{r}}{2} - \tan \frac{\tilde{t}-\tilde{r}}{2} \right). \qquad (1.60)$$

Our new coordinates $\tilde{t}, \tilde{r}$ extend only over a finite range of values, as can be seen from (1.59), (1.60). Thus, in the Penrose diagram of a Minkowski spacetime, see figure 1.7, we can depict the whole Minkowski spacetime, with a coordinate singularity along $\tilde{r} = 0$. All trajectories of uniformly moving particles (with velocity smaller than c) emerge from one single point, past infinity I^-, and all will eventually arrive at the one single point I^+, namely at future infinity. All incoming photons have their origin on the segment $\mathcal{I}^-$ (script I^- or 'scri minus'), lightlike past infinity, and will run into the coordinate singularity on the $\tilde{t}$-axis. All outgoing photons arise from the coordinate singularity and cease on the line $\mathcal{I}^+$, lightlike future infinity ('scri plus'). The entire spacelike infinity is mapped into the single point I^0.

Now, we have a really compact picture of the Minkowski space. Next, we would like to proceed along similar lines in order to obtain an analogy for the Schwarzschild spacetime.

1.6 Schwarzschild spacetime and the Penrose–Kruskal diagram

In relativity, light rays, the quasi-classical trajectories of photons, are null geodesics. In special relativity, this is quite obvious, since in Minkowski space the geodesics are straight lines and 'null' just means $v = c$. A more rigorous argument involves the solution of the Maxwell equations for the vacuum and the
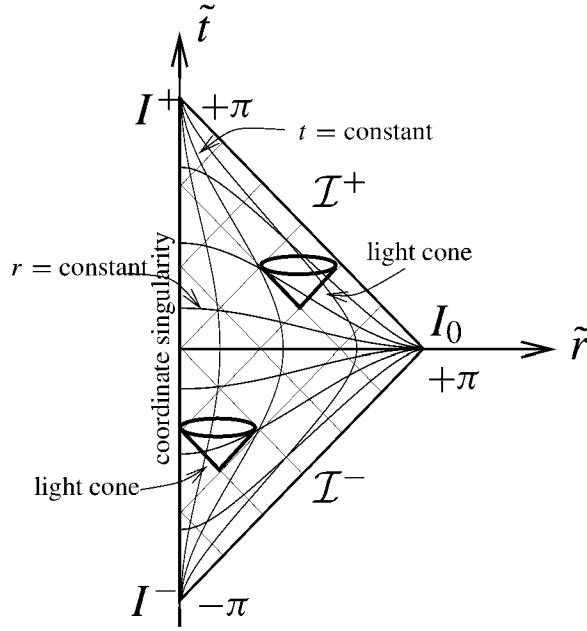
Figure 1.7. Penrose diagram of Minkowski spacetime.

subsequent determination of the normals to the wave surface (rays) which turn out to be null geodesics. This remains valid in general relativity. Null geodesics can be easily obtained by integrating the equation $0 = ds$. We find for the Schwarzschild metric, specializing to radial light rays with $d\phi = 0 = d\theta$, that

$$t = \pm \left(r + 2m \ln \left| \frac{r}{2m} - 1 \right| \right) + \text{constant}. \tag{1.61}$$

If we denote by r_0 the solution of the equation $r + 2m \ln |\frac{r}{2m} - 1| = 0$, we have for the t-coordinate of the light ray $t(r_0) =: v$. Hence, if $r = r_0$, we can use v to label light rays. In view of this, we introduce v and u[1]

$$v := t + r + 2m \ln \left| \frac{r}{2m} - 1 \right| \tag{1.62}$$

$$u := t - r - 2m \ln \left| \frac{r}{2m} - 1 \right|. \tag{1.63}$$

Then ingoing null geodesics are described by $v = $ constant, outgoing ones by $u = $ constant, see figure 1.8. We define *ingoing Eddington–Finkelstein coordinates* by replacing the 'Schwarzschild time' t by v. In these coordinates

---

[1] MTW [13] use capital letters, i.e. $u \to \tilde{U}$ and $v \to \tilde{V}$.
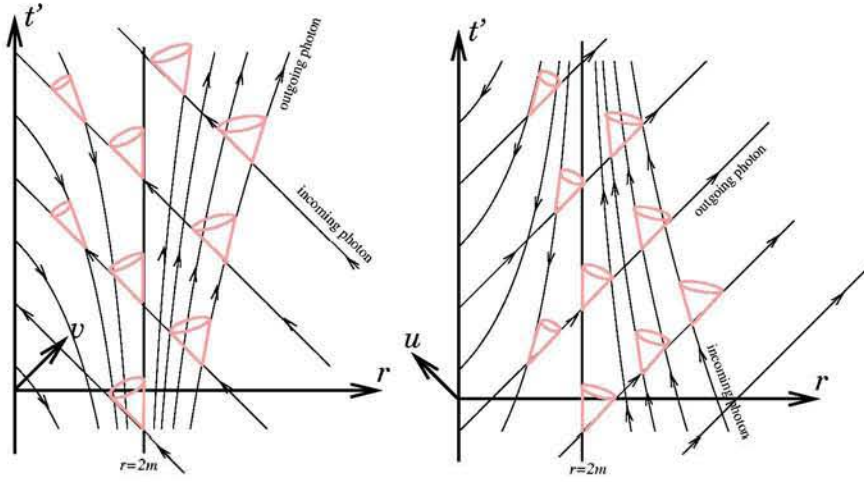
Figure 1.8. In- and outgoing Eddington–Finkelstein coordinates (where we introduce t' with $v = t' + r$, $u = t' - r$).

(v, r, θ, ϕ), the metric becomes

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dv^2 + 2dv\, dr + r^2\, d\Omega^2. \tag{1.64}$$

For radial null geodesics $ds^2 = d\theta = d\phi = 0$, we find two solutions of (1.64), namely $v = $ constant and $v = 4m \ln |r/2m - 1| + 2r + $ constant. The first one describes infalling photons, i.e. t increases if r approaches 0. At $r = 2m$, there is no longer any singular behavior for incoming photons. However, for outgoing photons, ingoing Eddington–Finkelstein coordinates are not well suited. Ingoing Eddington–Finkelstein coordinates are particularly useful for describing the gravitational collapse. Analogously, for outgoing null geodesics take (u, r, θ, ϕ) as the new coordinates. In these *outgoing Eddington–Finkelstein coordinates* the metric reads:

$$ds^2 = -\left(1 - \frac{2m}{r}\right) du^2 - 2\, du\, dr + r^2\, d\Omega^2. \tag{1.65}$$

Outgoing light rays are now described by $u = $ constant, ingoing light rays by $u = -(4m \ln |r/2m - 1| + 2r) + $ constant. In these coordinates, the hypersurface $r = 2m$ (the 'horizon') can be recognized as a null hypersurface (its normal is null or lightlike) and as a semi-permeable membrane.

Next we try to combine the advantages of in- and outgoing Eddington–Finkelstein coordinates in the hope of obtaining a fully regular coordinate system for the Schwarzschild spacetime. Therefore we assume coordinates (u, v, θ, ϕ).

Some (computer) algebra yields the corresponding representation of the metric:

$$ds^2 = -\left(1 - \frac{2m}{r(u, v)}\right) du\, dv + r^2(u, v)\, d\Omega^2. \tag{1.66}$$

Unfortunately, we still have a coordinate singularity at $r = 2m$. We can get rid of it by reparametrizing the surfaces $u = $ constant and $v = $ constant via

$$\tilde{v} = \exp\left(\frac{v}{4m}\right) \qquad \tilde{u} = -\exp\left(-\frac{u}{4m}\right). \tag{1.67}$$

In these coordinates, the metric reads ($r = r(\tilde{u}, \tilde{v})$ is implicitly given by (1.67) and (1.63), (1.62)):

$$ds^2 = -\frac{4r_S^3}{r(\tilde{u}, \tilde{v})} \exp\left(-\frac{r(\tilde{u}, \tilde{v})}{2m}\right) d\tilde{v}\, d\tilde{u} + r^2(\tilde{u}, \tilde{v})\, d\Omega^2. \tag{1.68}$$

Again, we go back from $\tilde{u}$ and $\tilde{v}$ to time- and spacelike coordinates:

$$\tilde{t} := \tfrac{1}{2}(\tilde{v} + \tilde{u}) \qquad \tilde{r} := \tfrac{1}{2}(\tilde{v} - \tilde{u}). \tag{1.69}$$

In terms of the original Schwarzschild coordinates we have[2]

$$\tilde{r} = \sqrt{\left|\frac{r}{2m} - 1\right|} \exp\left(\frac{r}{4m}\right) \cosh\frac{t}{4m} \tag{1.70}$$

$$\tilde{t} = \sqrt{\left|\frac{r}{2m} - 1\right|} \exp\left(\frac{r}{4m}\right) \sinh\frac{t}{4m}. \tag{1.71}$$

The Schwarzschild metric

$$ds^2 = \frac{4r_S^3}{r} \exp\left(-\frac{r}{2m}\right)(-d\tilde{t}^2 + d\tilde{r}^2) + r^2\, d\Omega^2 \qquad r = r(\tilde{t}, \tilde{r}) \tag{1.72}$$

in these *Kruskal–Szekeres* coordinates $(\tilde{t}, \tilde{r}, \theta, \phi)$, behaves regularly at the gravitational radius $r = 2m$. If we substitute (1.72) into the Einstein equation (via computer algebra), then we see that it is a solution of it for all $r > 0$. Equations (1.70), (1.71) yield

$$\tilde{r}^2 - \tilde{t}^2 = \left|\frac{r}{2m} - 1\right| \exp\left(\frac{r}{2m}\right). \tag{1.73}$$

Thus, the transformation is only valid for regions with $|\tilde{r}| > \tilde{t}$. However, we can find a set of transformations which cover the entire $(\tilde{t}, \tilde{r})$-space. They are valid in different domains, indicated here by I, II, III, and IV, to be explained later:

$$(\mathrm{I}) \begin{cases} \tilde{t} = \sqrt{\frac{r}{2m} - 1} \exp\left(\frac{r}{4m}\right) \sinh\frac{t}{4m} \\[2ex] \tilde{r} = \sqrt{\frac{r}{2m} - 1} \exp\left(\frac{r}{4m}\right) \cosh\frac{t}{4m} \end{cases} \tag{1.74}$$

[2] MTW [13] use v instead of $\tilde{t}$ and u instead of $\tilde{r}$.

$$\text{(II)} \begin{cases} \tilde{t} = \sqrt{1 - \dfrac{r}{2m}} \, \exp\left(\dfrac{r}{4m}\right) \cosh \dfrac{t}{4m} \\[4mm] \tilde{r} = \sqrt{1 - \dfrac{r}{2m}} \, \exp\left(\dfrac{r}{4m}\right) \sinh \dfrac{t}{4m} \end{cases} \tag{1.75}$$

$$\text{(III)} \begin{cases} \tilde{t} = -\sqrt{\dfrac{r}{2m} - 1} \, \exp\left(\dfrac{r}{4m}\right) \sinh \dfrac{t}{4m} \\[4mm] \tilde{r} = -\sqrt{\dfrac{r}{2m} - 1} \, \exp\left(\dfrac{r}{4m}\right) \cosh \dfrac{t}{4m} \end{cases} \tag{1.76}$$

$$\text{(IV)} \begin{cases} \tilde{t} = -\sqrt{1 - \dfrac{r}{2m}} \, \exp\left(\dfrac{r}{4m}\right) \cosh \dfrac{t}{4m} \\[4mm] \tilde{r} = -\sqrt{1 - \dfrac{r}{2m}} \, \exp\left(\dfrac{r}{4m}\right) \sinh \dfrac{t}{4m}. \end{cases} \tag{1.77}$$

The inverse transformation is given by

$$\left(\frac{r}{2m} - 1\right) \exp\left(\frac{r}{2m}\right) = \tilde{r}^2 - \tilde{t}^2 \tag{1.78}$$

$$\frac{t}{4m} = \begin{cases} \text{artanh}\, \tilde{t}/\tilde{r} & \text{for (I) and (III)} \\ \text{artanh}\, \tilde{r}/\tilde{t} & \text{for (II) and (IV).} \end{cases} \tag{1.79}$$

The Kruskal–Szekeres coordinates $(\tilde{t}, \tilde{r}, \theta, \phi)$ cover the entire spacetime (see figure 1.9). By means of the transformation equations we recognize that we need two Schwarzschild coordinate systems in order to cover the same domain. Regions (I) and (III) both correspond each to an asymptotically flat universe with $r > 2m$. Regions (II) and (IV) represent two regions with $r < 2m$. Since $\tilde{t}$ is a time coordinate, we see that the regions are time reversed with respect to each other. Within these regions, real physical singularities (corresponding to $r = 0$) move along the lines $\tilde{t}^2 - \tilde{r}^2 = 1$. From the form of the metric we can infer that the lightlike geodesics (and therewith the light cones $ds = 0$) are lines with slope $1/2$. This makes the discussion of the causal structure particularly simple.

Finally, we would like to represent the Schwarzschild spacetime in a manner analogous to the Penrose diagram of the Minkowski spacetime. To this end, we proceed along the same line as in the Minkowskian case. First, we again switch to null coordinates $v' = \tilde{t} + \tilde{r}$ and $u' = \tilde{t} - \tilde{r}$ and perform a conformal transformation which maps infinity into the finite (again, by means of the tangent function). Finally we return to a timelike coordinate $\hat{t}$ and a spacelike coordinate $\hat{r}$. We perform these transformations all in one go:

$$\tilde{t} + \tilde{r} = \tan \frac{\hat{t} + \hat{r}}{2} \tag{1.80}$$

$$\tilde{t} - \tilde{r} = \tan \frac{\hat{t} - \hat{r}}{2}. \tag{1.81}$$
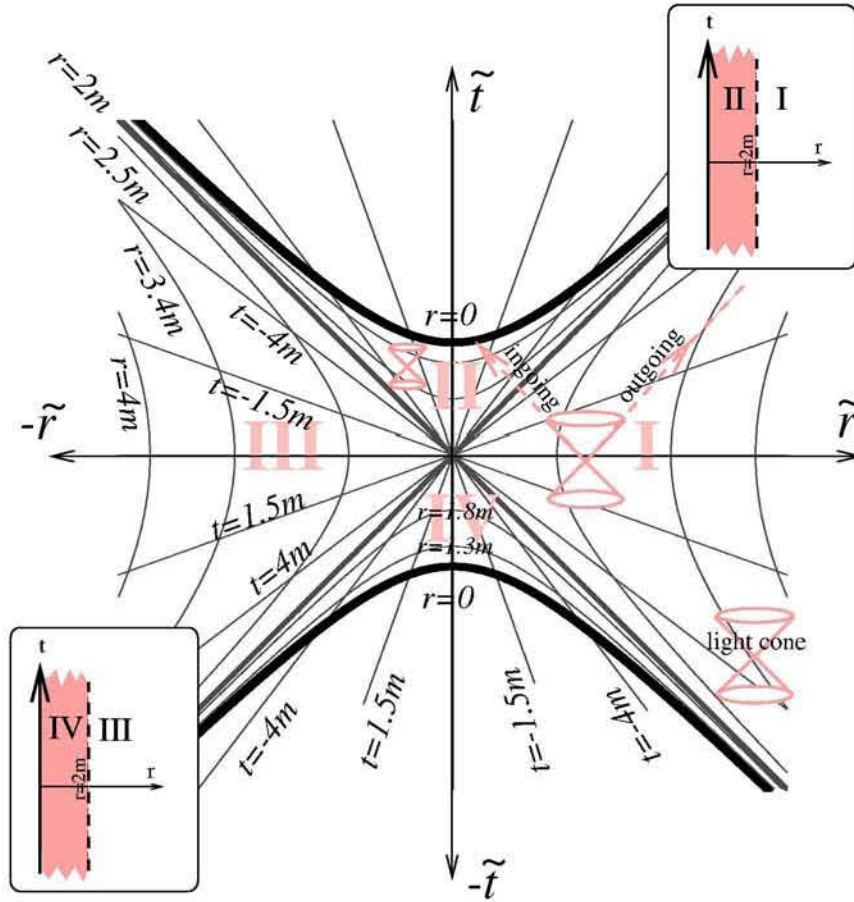
Figure 1.9. Kruskal-Szekeres diagram of the Schwarzschild spacetime.

The Schwarzschild metric then reads:

$$ds^2 = \frac{r_S^3}{r(\hat{r},\hat{t})} \frac{\exp\left(-\frac{r(\hat{r},\hat{t})}{2m}\right)(-d\hat{t}^2 + d\hat{r}^2)}{\cos^2 \frac{\hat{t}+\hat{r}}{2} \cos^2 \frac{\hat{t}-\hat{r}}{2}} + r^2(\hat{t},\hat{r})\, d\Omega^2 \qquad (1.82)$$

where the function $r(\hat{t}, \hat{r})$ is implicitly given by

$$\left(\frac{r}{2m} - 1\right)\exp\left(\frac{r}{2m}\right) = \tan\frac{\hat{t}+\hat{r}}{2} \tan\frac{\hat{t}-\hat{r}}{2}. \qquad (1.83)$$

The corresponding Penrose–Kruskal diagram is displayed in figure 1.10.
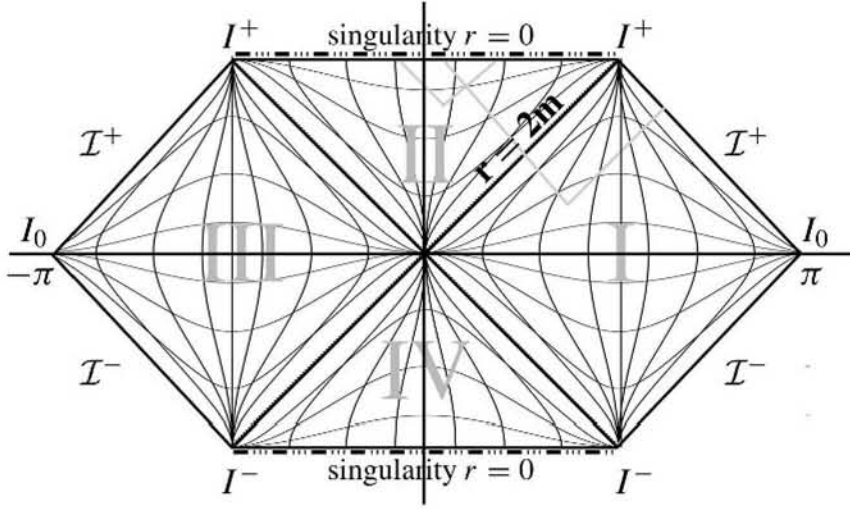
Figure 1.10. Penrose–Kruskal diagram of the Schwarzschild spacetime.

1.7 The interior Schwarzschild solution and the TOV equation

In the previous section we investigated the gravitational field outside a spherically symmetric mass distribution. Now it is time to have a look inside matter, see Adler *et al* [1]. Of course, in a first attempt, we have to make decisive simplifications on the internal structure of a star. We will consider cold catalyzed stellar material during the later phase of its evolution which can be reasonably approximated by a perfect fluid. The typical mass densities are in the range of $\approx 10^7$ g cm^{-3} (white dwarfs) or $\approx 10^{14}$ g cm^{-3} (neutron stars, i.e. pulsars). In this context we assume vanishing angular momentum.

We start again from a static and spherically symmetric metric

$$ds^2 = -e^{A(r)}c^2\,dt^2 + e^{B(r)}\,dr^2 + r^2\,d\Omega^2 \qquad (1.84)$$

and the energy–momentum tensor

$$T_{\mu\nu} = \left(\rho + \frac{p}{c^2}\right)u_\mu u_\nu + p g_{\mu\nu} \qquad (1.85)$$

where $\rho = \rho(r)$ is the spherically symmetric mass density and $p = p(r)$ the pressure (isotropic stress). This has to be supplemented by the equation of state which, for a simple fluid, has the form $p = p(\rho)$.

We compute the non-vanishing components of the field equation by means of computer algebra as (here $()' = d/dr$)

$$-e^B \kappa r^2 c^2 \rho + e^B + B'r - 1 = 0 \qquad (1.86)$$

$$-e^B \kappa p r^2 - e^B + A'r + 1 = 0 \qquad (1.87)$$

$$-4e^B \kappa p r + 2A''r + (A')^2 r - A'B'r + 2A' - 2B' = 0. \qquad (1.88)$$

The (ϕ, ϕ)-component turns out to be equivalent to the (θ, θ)-component. For convenience, we define a *mass function* $m(r)$ according to

$$e^{-B} =: 1 - \frac{2m(r)}{r}. \qquad (1.89)$$

We can differentiate (1.89) with respect to r and find, after substituting (1.86), a differential equation for $m(r)$ which can be integrated, provided $\rho(r)$ is assumed to be known:

$$m(r) = \int_0^r \frac{\kappa}{2} \rho(\xi) c^2 \xi^2 \, d\xi. \qquad (1.90)$$

Differentiating (1.87) and using all three components of the field equation, we obtain a differential equation for A:

$$A' = -\frac{2p'}{p + \rho c^2}. \qquad (1.91)$$

We can derive an alternative representation of A' by substituting (1.89) into (1.87). Then, together with (1.91), we arrive at the *Tolman–Oppenheimer–Volkoff* (TOV) equation

$$p' = -\frac{(\rho c^2 + p)(\mathbf{m} + \kappa p r^3/2)}{\mathbf{r}(\mathbf{r} - 2m)}. \qquad (1.92)$$

The Newtonian terms are denoted by boldface letters. The system of equations consisting of (1.90), (1.91), the TOV equation (1.92), and the equation of state $p = p(\rho)$ forms a complete set of equations for the unknown functions $A(r)$, $\rho(r)$, $p(r)$, and $m(r)$, with

$$ds^2 = -e^{A(r)} c^2 \, dt^2 + \frac{dr^2}{1 - \frac{2m(r)}{r}} + r^2 \, d\Omega^2. \qquad (1.93)$$

These differential equations have to be supplemented by initial conditions.

In the center of the star, there is, of course, no enclosed mass. Hence we demand $m(0) = 0$. The density has to be finite at the origin, i.e. $\rho(0) = \rho_c$, where ρ_c is the density of the central region. At the surfaces of the star, at $r = R_\odot$, we have to match matter with vacuum. In vacuum, there is no pressure which requires $p(R_\odot) = 0$. Moreover, the mass function should then yield the total mass of the star, $m(R_\odot) = M$. Finally, we have to match the components of the metric. Therefore, we have to demand $\exp[A(R_\odot)] = 1 - 2m(R_\odot)/R_\odot$.

Equations (1.86), (1.87), (1.88) and certain regularity conditions which generalize our boundary conditions, i.e.

- the regularity of the geometry at the origin,
- the finiteness of the central pressure and density,
- the positivity of the central pressure and density,
- the positivity of the pressure and density,
- the monotonic decrease in pressure and density,

impose conditions on the functions ρ and p. Then, even without explicit knowledge of the equation of state, the general form of the metric can be determined. For most recent work, see Rahman and Visser [16] and the literature given there.

We can obtain a simple solution, if we assume a constant mass density

$$\rho = \rho(r) = \text{constant.} \tag{1.94}$$

One should mention here that ρ is not the physically observable fluid density, which results from an appropriate projection of the energy–momentum tensor into the reference frame of an observer. Thus, this model is not as unphysical as it may look at first. However, there are serious but more subtle objections which we will not discuss further in this context.

When $\rho = $ constant, we can immediately integrate (1.89) and thus obtain the metric component $\exp(B)$. Also (1.91) can be integrated. Then, after some more elementary integrations, we can make use of the boundary conditions. Finally, we arrive at the *interior and exterior Schwarzschild solution* for a spherically symmetric body [20]:

$$ds^2 = \begin{cases} -\left(\dfrac{3}{2}\sqrt{1 - \dfrac{R_\odot^2}{\hat{R}^2}} - \dfrac{1}{2}\sqrt{1 - \dfrac{r^2}{\hat{R}^2}} \right)^2 c^2\, dt^2 \\[4mm] \quad + \dfrac{1}{1 - \frac{r^2}{\hat{R}^2}}\, dr^2 + r^2\, d\Omega^2 & r \le R_\odot \\[6mm] -\left(1 - \dfrac{2m}{r} \right) c^2\, dt^2 + \dfrac{1}{1 - \frac{2m}{r}}\, dr^2 + r^2\, d\Omega^2 & r > R_\odot \end{cases} \tag{1.95}$$

with

$$\hat{R} := \sqrt{\frac{3}{\kappa \rho c^2}} \qquad \rho = \text{constant.} \tag{1.96}$$

For the sun we have $M_\odot \approx 2 \times 10^{30}$ kg, $R_\odot \approx 7 \times 10^8$ m and subsequently $\rho_\odot \approx 1.4 \times 10^3$ kg m^{-3}. This leads to $\hat{R} \approx 3 \times 10^{11}$, i.e. the radius of the star $R_\odot$ is much smaller than $\hat{R}$: $R_\odot < \hat{R}$. Hence the square roots in (1.95) remain real.

Visualization and comparison with a 'Newtonian' star

From the continuous matching of the g_{rr}-component we can derive the relation $1 - 2m/R_\odot = 1 - R_\odot^2/\hat{R}^2$. Together with the definition of the Schwarzschild

radius we find for the total gravitating mass of the star

$$M = \frac{4\pi}{3} R_\odot^3 \rho.$$

(1.97)

Another method for obtaining the total mass is to multiply the density ρ by the spatial volume of the star at a given time t_0. However, the total mass calculated that way is *larger* than the total gravitating mass (1.97). This is due to the fact that not mass (that is 'rest mass') alone but mass-energy gravitates. The negative gravitational binding forces thus contribute to the gravitating mass which appears in the metric.

Finally, some words about the geometry of the Schwarzschild spacetime. We can visualize its structure by means of an embedding in the following way: In the equatorial plane $\vartheta = \pi/2$ at a prescribed time $t = t_0$, the metric reads $(\hat{R}^2 = R_\odot^3/2m)$:

$$ds^2 = \begin{cases} \left(1 - \dfrac{2mr^2}{R_\odot^3}\right)^{-1} dr^2 + r^2 \, d\varphi^2 & \text{for } r \leq R_\odot \\[4mm] \left(1 - \dfrac{2m}{r}\right)^{-1} dr^2 + r^2 \, d\varphi^2 & \text{for } r > R_\odot. \end{cases}$$

(1.98)

These metrics are equivalent to 2D metrics induced by the 3D *Euclidean* metric on a sphere or a hyperboloid, respectively. The 3D Euclidean metric is $ds^2 = dr^2 + r^2 \, d\phi^2 + dz^2$. A surface rotationally symmetric around the z-axis is described by a parametrization $z = z(r)$. The metric induced on this surface is $ds^2 = [1 + (dz/dr)^2] \, dr^2 + r^2 \, d\phi^2$. By comparison with the previous metrics, we extract differential equations for $z(r)$ which can be easily solved. At $r = R_\odot$, the surfaces are continuously joined (see figure 1.11).

Outside, we have the usual vacuum Schwarzschild geometry which was discussed extensively in the previous section. We may add a few remarks. Obviously, a circle (or sphere, respectively) around the origin has a circumference of $2\pi r$, where r is the radial Schwarzschild coordinate. We also observe that the proper distance measured by a freely falling observer (who, in our picture, moves radially on the hyperboloid) is larger than the coordinate distance Δr. Inside the star we have the three-geometry of a sphere with radius $\hat{R}$. Far away from the star we find flat Euclidean geometry.

The structure of this three-geometry resembles the Newtonian case. Inside, we have a conformally flat space, where the Weyl ('trace-free part of the curvature') vanishes and the Ricci tensor is proportional to the mass-energy density. In the Newtonian case, the trace of the tidal matrix (the analogy to curvature) is proportional to the mass density, and, subsequently, its trace-free part vanishes. Outside, in vacuum, it is the other way around. There the trace parts are zero ($K_{aa} = 0$ and $\mathrm{Ric}_{\alpha\beta} = 0 = R$). The Newtonian tidal acceleration

Z (additional) embedding space dimension

x

exterior

interior

r

Δr_2

Δr_1

y
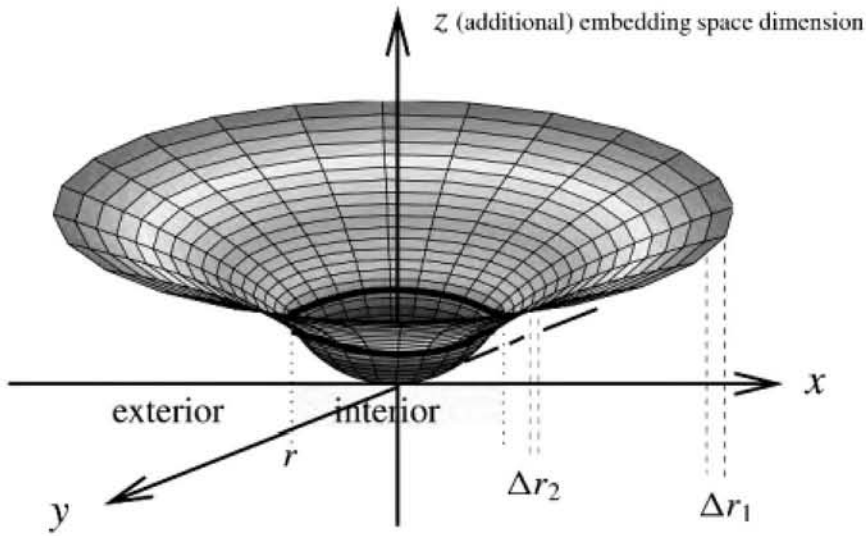
Figure 1.11. Geometry of Schwarzschild spacetime.

matrix is trace free and reads (in Cartesian coordinates, assume $\mathbf{r} = (0, 0, r)$):

$$K_{ab} \stackrel{*}{=} \frac{GM}{r^3} \text{diag}(1, 1, -2). \tag{1.99}$$

In Einstein's theory we have to use the equation for the geodesic deviation (1.39) in order to calculate the relative acceleration of two freely falling test particles. For the comoving observer, with $u^\alpha = (c, 0, 0, 0)$ and in an orthonormal frame, we find

$$\ddot{v}^\mu \stackrel{*}{=} c^2 R^\mu{}_{0\nu 0} v^\nu = \frac{GM}{r^3} \text{diag}(1, 1, -2)(v^1, v^2, v^3). \tag{1.100}$$

Thus, in a special frame, we have the same tidal accelerations as in the Newtonian case.

Accordingly, the gravitational field of a spherically symmetric body in Newton's 3D theory is very naturally embedded into Einstein's 4D theory.

1.8 Computer algebra

As a typical example, we will demonstrate how to obtain the exterior Schwarzschild solution by means of the computer algebra system Reduce and its package Excalc for applications in differential geometry. When Reduce is called, it prompts the user for input. Each statement has to be terminated by a semicolon (output is displayed) or by a dollar sign (output is suppressed). After pressing

the return-key, the computer (hopefully) will produce a result. So let us start by loading the package Excalc:

```
load excalc ;
```

Then we define the metric. Therefore we first introduce the functions ν and λ which enter the ansatz for the metric

```
pform{nu,lamb}=0;
```

and declare which variables they depend on:

```
fdomain nu = nu(r,t), lamb=lamb(r,t) ;
```

Subsequently, we define coframe and metric:

```
coframe   o(t) = d t ,
          o(r) = d r ,
          o(theta) = d theta ,
          o(phi) = d phi
with metric g =   - exp(nu) * o(t) * o(t)
                  + exp(lamb) * o(r) * o(r)
                  + r**2 * o(theta) * o(theta)
                  + r**2 *sin(theta)**2 * o(phi) * o(phi) ;
```

Excalc is a package designed to perform calculations with exterior differential forms. It is convenient to compute partial derivatives of scalar functions as follows.

$$\partial_\alpha \phi = e_\alpha \rfloor \, d\phi = e_\alpha \rfloor (\partial_\alpha \phi \, dx^\alpha) \qquad (1.101)$$

where we have introduced the vector basis e_α dual to the coframe, i.e. $e_\alpha \rfloor \, dx^\beta = \delta_\beta^\alpha$ ($\rfloor$ is the interior product sign). Accordingly, we compute the vector basis

```
frame e ;
```

and define the Christoffel symbol

```
pform chris(i,j,k) = 0 ;
index_symmetries chris(i,j,k): symmetric in {j,k}            ;

chris(-i,-j,-k) := (1/2) * (   e(-k)_|d g(-i,-j)
                             + e(-j)_|d g(-i,-k)
                             - e(-i)_|d g(-j,-k) );
```

Next, we compute the Riemannian curvature tensor. By means of the declaration `index_symmetries`, we can explicitly implement the index symmetries of tensors, which saves a lot of memory and computation time. Moreover, the printed output then encompasses automatically only independent components.

```
pform riem(i,j,k,l) = 0;
index_symmetries riem(i,j,k,l): antisymmetric in {i,j},{k,l}
```

```
                            symmetric in {{i,j},{k,l}} ;

riem(i,-j,-k,-l) :=   e(-k)_|d chris(i,-j,-l)
                    - e(-l)_|d chris(i,-j,-k)
                    + chris(i,-m,-k) * chris(m,-j,-l)
                    - chris(i,-m,-l) * chris(m,-j,-k) ;
```

Then, we introduce the Ricci tensor, curvature scalar, and Einstein tensor.

```
pform ricci(i,j)=0 ;
ricci(-i,-j) := riem(k,-i,-k,-j) ;

pform rscalar = 0;
rscalar := ricci(-i,i) ;

pform einstein(i,j) = 0 ;
einstein(i,j) := ricci(i,j) - (1/2) * rscalar * g(i,j) ;
```

Now we implement the vacuum field equation:

```
pform zero(i,j) = 0 ;
zero(i,j) :=  einstein(i,j) + kosmo * g(i,j) ;
```

The next step is to look at the output and to get some ideas of how to proceed...With a computer algebra system, we can very easily manipulate systems of equations in order to obtain new, simpler equations. By entering (num yields the numerator of a fraction):

```
0 = num(zero(t,t)) + num(zero(r,r)) ;
```

we get

$$0 = \partial_r \lambda r + \partial_r \nu r. \tag{1.102}$$

Accordingly, the sum $f := \lambda + \nu$ has to be independent of r and thus is a function of t alone. Then we can perform a rescaling of the time coordinate

$$t \longrightarrow t' = \int dt \, e^{f(t)/2} \tag{1.103}$$

such that

$$dt' = e^{f(t)/2} dt. \tag{1.104}$$

Hence, the ansatz for the metric does not change, apart from the (t, t)-component

$$e^{\nu(r,t)} dt^2 = e^{\nu(r,t') - f(t')} dt'^2 =: e^{\nu'} dt'^2 \tag{1.105}$$

or

$$\nu = \nu' + f(t). \tag{1.106}$$

Thus,

$$\lambda = -\nu'. \tag{1.107}$$

Eventually, we can put

```
lamb := - nu ;
```

and suppress the prime from now on. Next, we note that

$$0 = \texttt{zero(r, t)} = \frac{\partial_t \nu}{r}. \tag{1.108}$$

Consequently, the function ν cannot depend on t. We take this into account with

```
@(nu,t) := 0 ;
```

For convenience, we get rid of the exp-functions:

```
pform psi = 0 ;
fdomain psi = psi(r) ;
nu := log(psi) ;

zero(i,j) := zero(i,j);
```

The (r, r)-component of the field equation can be solved for $\partial_r \psi$. We can do this with the computer by means of the `solve` operator

```
solve(zero(r,r)=0,@(psi,r)) ;
```

We then substitute the result into the field equation

```
@(psi,r) := ( - psi + 1 - kosmo*r**2)/r;
```

It turns out that all components of the field equation are already fulfilled. The task of solving the ordinary differential equation remains:

$$\partial_r \psi + \frac{\psi}{r} - \frac{1}{r} + \Lambda r = 0. \tag{1.109}$$

This may be done by means of an appropriate package, like the Reduce package `odesolve`.

```
load odesolve ;
odesolve(df(psi,r)-@(psi,r),psi,r) ;
```

By setting the integration constant to $-2m$, we finally arrive at

$$\psi = 1 - \frac{2m}{r} - \frac{\Lambda}{3}r^2. \tag{1.110}$$

References

[1] Adler R, Bazin M and Schiffer M 1975 *Introduction to General Relativity* 2nd edn (New York: McGraw-Hill)

[2] Ciufolini I, Pavlis E, Chieppa F, Fernandes-Vieria E and Pérez-Mercader J 1998 Test of general relativity and the measurement of the Lense–Thirring effect with two earth satellites *Science* **279** 2100–4

[3] Ciufolini I and Wheeler J A 1995 *Gravitation and Inertia* (Princeton, NJ: Princeton University Press)

[4] Einstein A 1992 *The Meaning of Relativity* (Princeton, NJ: Princeton University Press)

[5] Einstein A 1914 Zum Gegenwärtigen Stande des Gravitationsproblems *Phys. Zeit.* **14** 1249–66

[6] Everitt C W F *et al* Gravity Probe B: Countdown to Launch in [12], pp 52–82

[7] de Felice F and Clarke C J S 1990 *Relativity on Curved Manifolds* (Cambridge: Cambridge University Press)

[8] Flowers J L and Petley B W 2001 Progress in our knowledge of the fundamental constants in physics *Rep. Prog. Phys.* **64** 1191–246

[9] Frolov V P and Novikov I D 1998 *Black Hole Physics. Basic Concepts and New Developments* (Dordrecht: Kluwer)

[10] Hawking S W and Ellis G F R 1973 *The Large Scale Structure of Spacetime* (Cambridge: Cambridge University Press)

[11] Heinicke C 2001 The Einstein 3-form and its equivalent 1-form L_α in Riemann–Cartan space *Gen. Rel. Grav.* **33** 1115–31

[12] Lämmerzahl C, Everitt C W F and Hehl F W (eds) 2001 *Gyros, Clocks, Interferometers...: Testing Relativistic Gravity in Space (Lecture Notes in Physics 562)* (Berlin: Springer)

[13] Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco: Freeman)

[14] Newton I 1999 *The Principia: Mathematical Principles of Natural Philosophy* translation by B I Cohen, A Whitman, and J Budenz; preceded by *A Guide to Newton's Principia* by B Cohen (Berkeley, CA: Univeristy of California Press)

[15] Poincaré J H 1952 *Science and Hypothesis* translation from the French (New York: Dover)

[16] Rahman S and Visser M 2002 Spacetime geometry of static fluid spheres *Class. Quant. Grav.* **19** 935–52. See also http://www.arXiv.org/abs/gr-qc/0103065

[17] Rindler W 2001 *Relativity. Special, General, and Cosmological* (Oxford: Oxford University Press)

[18] Schwarzschild K 1998 *The Collected Papers of Albert Einstein. Vol. 8, The Berlin Years: Correspondence, 1914–1918* ed R Schulmann, M Janssen and J Illy (Princeton, NJ: Princeton University Press)

[19] Schwarzschild K 1916 Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie *Sitzungsber. Preuss. Akad. Wiss. Berlin* 189–96

[20] Schwarzschild K 1916 Über das Gravitationsfeld einer Kugel aus inkompressibler Flüssigkeit nach der Einsteinschen Theorie *Sitzungsber. Preuss. Akad. Wiss. Berlin* 424–34

[21] Will C M 2001 The confrontation between general relativity and experiment *Living Rev. Relativity* **4** 4 [Online article]: cited on 17 Nov 2001. http://www. livingreviews.org/Articles/Volume4/2001-4will/

Chapter 2

The Milky Way: structure, constituents and evolution

Susanne Hüttemeister
Ruhr-Universität Bochum, Germany

After having learned, in chapter 1, about an idealized relativistic star, we now turn to the Milky Way which consists of about 10^{11} stars and has a diameter of about 10^5 light years. The historical and current developments in our understanding of the Milky Way are described, based mainly on observations which are becoming increasingly accurate. The evolution and formation of galaxies in a cosmological context is also discussed as well as the relation of their properties to their central black hole. The black hole at the center of the Milky Way—among all galaxies the best studied case—will be the focus of our considerations in subsequent chapters.

2.1 The overall structure of the Milky Way

2.1.1 Deducing the large-scale structure of the Galaxy

When viewed on a moonless night from a place far away from the pollution of city light, the 'band' of the Milky Way stands out clearly and in great detail. With binoculars, a small telescope or even the naked eye we may discern not only numerous individual stars, but also star clusters, gas nebulae shining red in the light of ionized hydrogen (H II regions), dust lanes, some of which break up into fine filaments, and dark clouds. Comparing the view from the northern and southern hemispheres, we find that the density of stars, clusters but also that of dark clouds is highest toward the constellations Sagittarius and Scorpius. With our present-day knowledge about Galactic structure, it is quite possible to visualize our place within the disk of the Milky Way and even to get some feeling of being located within a three-dimensional structure. However, it is also clear that we are at some disadvantage when trying to deduce the true structure of the Milky

Way, due to our viewpoint inside it, resulting in an 'edge-on' viewing geometry imposed by the location of the Sun within the Galactic Disk.

Indeed, while our understanding of the general appearance of the Milky Way has made great progress over the last century, progress which has become—as for almost all other astronomical phenomena—ever faster in recent decades, there are still considerable blanks to fill in. The Galactic Center itself as well as the far side of the Milky Way, beyond the center, are almost entirely inaccessible to optical astronomy, making research on the Galactic Center region and, to some degree, on large-scale Galactic structure a domain for other wavelength ranges. Of these, only the radio and part of the infrared domain are accessible to ground-based telescopes, and even for these, much of the necessary technology has only been developed recently.

Sometimes, it seems easier to discern the structure of external galaxies than that of our own Milky Way, at least on a large scale that does not require high resolution. A number of external galaxies have indeed been put forward as 'templates' for the Milky Way. NGC 891 or NGC 4565 may present us with an edge-on view of the Milky Way (e.g. van der Kruit 1984), while NGC 1232 (Möllenhoff *et al* 1999, see figure 2.1) may look similar to our own galaxy when seen face-on. Several components or constituents (which we will examine more closely in later sections) are seen readily in these external galaxies: in face-on views, spiral arms, inter-arm regions and integral color changes when moving in from the disk to the central region are obvious, while edge-on views show the thinness of the disk, with an even thinner dust lane cutting through its central part, as well as the oval bulge region in the center.

However, studying external galaxies does not really solve the problem of the structure of our Galaxy: there are many types of galaxy, presumably in many evolutionary stages, and even within one class, e.g. spiral galaxies, the members show a great variety of more or less obvious differences. In fact, no two galaxies are exactly alike. Thus, using external galaxies as Milky Way templates requires much knowledge about the structure of the Milky Way just to select the right galaxies for companion. Ideally, high-resolution studies of the constituents of the Milky Way and, by necessity, lower-resolution work on (many) external galaxies should and can complement each other, shedding light on the more general problems of galactic structure and evolution. However, to solve specific questions on the structure of the Milky Way as a unique object, we have little choice but to turn to our Galaxy itself and try to overcome the problems due to our position within it.

2.1.2 Unveiling Galactic structure: history

The awareness that the main constituents of the Milky Way are stars came with the invention of the telescope. Galilei stated in 1610 that '—the Galaxy is nothing else but a mass of innumerable stars planted together in clusters—' (quoted from Weaver 1975; the material in this section is largely adapted from his articles

Figure 2.1. The famous ESO VLT image of NGC 1232, a possibly Milky Way template galaxy.

(Weaver 1975a, b) and Hoskin (1985)). The shape and configuration of the stars proved, however, difficult to determine.

In 1750, Wright published what is widely regarded as the first disklike picture of the Milky Way. A plate in his 'An Original Theory or New Hypothesis of the Universe' depicts what appears to be a stellar disk, with the sun within it and lines-of-sight drawn across that clearly and correctly explain why the Milky Way is perceived as a bright band—we see many more stars when we look along a line-of-sight within the disk than when we look perpendicularly to the plane of the disk. However, Wright's view only seems 'modern': he thought of the 'disk' as part of a very large spherical shell, with a radius so large that the curvature was hardly perceptible. In the center of the sphere was 'Heaven, the Abode of God', while Wright assumed the far side to be 'the Shades of Darkness and Dispare, the Desolate Regions of ye Damned'. This certainly appears to be a rather unconventional view of both the Galactic Center and extragalactic space from

a modern perspective, but Wright was indeed the first to assume that the Sun was rotating around some central object. However, Wright, and his contemporaries, while trying to develop world views consistent with observations, did so with philosophical or theological reasoning rather than experiments. Thus, the insights they arrived at belong to the realm of natural philosophy more than empirical science.

Kant knew of Wright's ideas, and took them one step further. In his 'Allgemeine Naturgeschichte und Theorie des Himmels' (1755), he arrived at a 'true' disk picture, extending the hierarchical structure by deducing the existence of 'Welteninseln' or 'Island Universes'—external galaxies.

The viewpoint of natural philosophy was changed fundamentally to one of empirical science when W Herschel, telescope builder and tireless observer, introduced not only observational but also statistical methods into the study of Galactic structure. He and his sister performed star counts for many lines-of-sight along a great circle, and, inventing the methodology of stellar statistics for this purpose, arrived at a picture of the Milky Way as a flattened, irregularly shaped object with the Sun close to the center. Herschel had to make two assumptions both of which later proved to be incorrect: he assumed an identical space density for the stars everywhere within the Milky Way and that stellar brightness was roughly indicative of distance, allowing his telescope to reach the edge of the system.

Herschel himself realized later in life that these assumptions were flawed, but his and other models based on this method were reproduced and refined for many decades after his death. In 1922, Kapteyn published a summary of all these efforts: his 'Kapteyn Universe' showed a circular, lens-shaped galaxy about 15 kpc in size, with the Sun again close to the center.

Even before Kapteyn's model was published, it was under what proved later to be a decisive attack. In 1915, Shapley had started to pin down the location of globular clusters, the distances of which he could determine by a method based on variable stars (Shapley 1918). Globular clusters are (today) known to be old halo objects and their distribution is not confined to the disk of the Galaxy. Thus they can be seen optically at very large distances. Shapley found that the center of the globular cluster system was located outside the Kapteyn Milky Way, a situation that is dynamically impossible. Thus, he arrived at a picture of a Galaxy that was much larger than before (in fact, too large), with the Sun relegated to a position closer to the edge than to the center.

Uncertainty about the nature of the spiral nebulae added to a confusing situation: while most of the supporters of a Kapteyn-like universe believed that nebulae like M 31 or M 51 were galaxies similar to our own, from the beginning of measurements of radial motions and the detection of novae in spirals, partisans of Shapley's view held to the opinion that spiral nebulae were part of the larger Milky Way. The 'new star' S Andromedae in M 31 seemed to support their view, being inexplicably bright if located in another galaxy. Another complication was the alleged measurement of rotation in spirals on photographic plates, by van

Maanen, known to be a meticulous observer. While van Maanen's error was never quite explained, S Andromedae turned out to be a supernova, intrinsically much brighter than any 'new star' observed before.

Shapley and Curtis, an adherent of a small (Kapteyn) Galaxy and the extragalactic nature of spirals, met in a 'Great Debate' in Washington in 1920, where arguments were exchanged, and no resolution reached. In hindsight, we know that the views of both groups were partly true: Shapley's assessment of the size of the Milky Way and the Sun's location was close to being correct, while the spiral nebulae are indeed external galaxies.

This question was settled only a few years later, when Hubble found Cepheid variables in spiral nebulae, determining their distances, and—a little later—their general recession, which later became famous as the 'Hubble law' of galaxy redshift. Oort was the first to analyze Galactic rotation in 1927, finding a position for the center that roughly agreed with Shapley's determination.

Finally, in 1930 the reason for much of the disagreement and confusion became clear when Trümpler demonstrated the existence of Galactic extinction in his investigations of photometric distances, linear scales and reddening of Galactic open clusters. Thus, an absorbing dust component for the interstellar medium (ISM) was established as an important constituent of the Milky Way, even in regions where its presence was not obvious as dark clouds or filaments. These were only now realized to be absorbing layers of material, and not starless voids or 'holes in the sky' (as the astrophotography pioneer E E Barnard thought). It became clear that in the optical wavelength range our view is limited to a few kpc, and the Galactic structure at large cannot be inferred from star counts (though the local disk structure can still be investigated by stellar statistics).

For this reason the center of the Milky Way, as officially adopted by the International Astronomical Union in 1959 as the origin for the Galactic coordinate system, was eventually based on the detection of strong radio emission from the nucleus of our Galaxy (Piddington and Minnett 1951).

2.1.3 'External' views

More than seven decades later, data from many wavelength ranges, many of which penetrate the layer of dust extinction, are at our disposal to derive Galactic structure. Still, all our direct observational views are (and will remain for the indefinite future) internal and edge-on, with all the associated problems. By now, they cover the entire electromagnetic spectrum. This includes the radio regime, where we encounter non-thermal synchrotron emission from relativistic electrons at long cm wavelengths, emission from neutral atomic hydrogen at 21 cm and molecules, most prominently CO, which are used to trace the molecular gas component, at mm wavelengths. The far and mid-infrared region is dominated by thermal dust emission, while in the near infrared (NIR) we encounter emission from low mass, cool stars. The optical and UV bands are most affected by interstellar extinction, limiting our view to nearby stars,

Draft of the 'Berlin Milky Way'

Figure 2.2. Draft of the 'external view' of the Milky Way developed by the author and E Janssen for the exhibition 'Seven Hills' in Berlin (2000).

dust clouds and H II regions. X-rays, only accessible to satellite observatories, provide information on the hot component of the ISM, while the highest energy γ-rays mostly arise in collisions of cosmic rays with hydrogen atoms (see nvo.gsfc.nasa.gov/mw/milkyway.html for multiwavelength views of the Galaxy).

It is our task to piece together a coherent picture of Galactic structure based on these diverse sources of information on the different constituents of the Galaxy. A view of the Milky Way as it might appear to an external observer is necessarily an artist's conception, and partially based on (hopefully) educated guesses. Two attempts have been made: J Lomberg's painting at the National Air and Space Museum in Washington and a view of the Milky Way (figure 2.2) put together by the author and the artist E Janssen (European Southern Observatory) for the exhibition 'Seven Hills—Images and Perspectives for the 21st Century' in Berlin

(2000). Both images are based on the available data. Lomberg's view focused on the Sun and its fairly well-known surroundings, thus the unknown details on the far side of the Galaxy are suitably blurred with distance. The 'Berlin' Milky Way is shown face-on, thus we had to invent a likely structure on the far side; we settled on an overall shape for the spiral structure, for which we chose a model with four main spiral arms, in accordance with most (but not all) the evidence.

We will examine the main constituents in more detail later, but will familiarize ourselves with the main components of the large-scale Galactic structure at this point: external views are dominated by the distinct constituents of the disk and specifically the spiral arms: young massive blue stars, reddish H II regions as the sites of stars still in the process of ionizing and dissolving their birth cloud, young clusters and dust lanes and filaments, the cradles of ongoing star formation. The disk has a diameter $\leq$ 30 kpc (but no sharp edge), a thickness of at most 1 kpc (depending on population), and its surface brightness falls roughly exponentially, with a scale length of 2.2–2.8 kpc. The Sun is at a distance of $\sim$8 kpc (7–8.5 kpc) from the center. In the vicinity of the Sun, the surface mass density is $\Sigma_{tot} = (71 \pm 6)\ M_\odot\ pc^{-2}$, a fairly certain value confirmed by a number of investigations (Kuijken and Gilmore 1991, Olling and Merrifield 2001). The volume density in the disk is far less certain: values range from $\rho_{tot} = 0.11$ to 0.076 $M_\odot\ pc^{-3}$ (Crezé *et al* 1998, Holmberg and Flynn 2000). The stellar surface density close to the Solar circle is in the range $\Sigma_* = (25–50)\ M_\odot\ pc^{-2}$. Correspondingly, the local surface density of dark matter is only very poorly known; it is estimated at $\Sigma_{DM} = (10–35)\ M_\odot\ pc^{-2}$. There is some indication of a stellar warp in the disk, and stronger evidence for a warped distribution of dust and gas.

A weak bar (or triaxial bulge) of diameter $\sim$3 kpc is seen in the inner part of the Galaxy, and the general color changes from whitish blue in the disk region to orange or reddish in the bulge, indicative of a change in stellar population from a mix dominated, at least in luminosity, by young stars to a population made up mostly of older stars.

Beyond and above the Galactic disk extends the tenuous halo with its scattered old stars and globular clusters, the density of which is, however, strongly concentrated toward the center, and which, in addition, shows substructure and subpopulations (e.g. Zinn 1985, Burkert and Smith 1997). Dark matter, likely to be a very important halo component and decisive in structure formation scenarios, remains enigmatic, even though it is thought to dominate the overall mass budget of the Galaxy.

The Sun is located within a structure that is sometimes called the 'Orion Arm', but would better be named the 'Orion Spur'. Evidence points to it not being a real spiral arm, but a short protrusion, as seen frequently in images of external galaxies. We may be fortunate not to be positioned in the midst of a 'real' spiral arm: while spiral galaxies, seen face-on, are largely transparent (Xilouris *et al* 1999, Bosma *et al* 1992), the enhanced opacity caused by local dust clouds likely to be encountered in a strong spiral arm might have limited the region of the Milky

Way accessible to optical studies still further. It might even have rendered many or all external galaxies invisible in the optical wavelength range, greatly expanding the galactic 'zone of avoidance' and delaying the development of extragalactic astronomy to a time where measurements of extinction-free tracers were possible, thus profoundly changing the history of our view of the cosmos and the Galaxy retold briefly in previous sections.

2.2   The constituents

In this section, we will examine the constituents of the Galaxy and try to assemble a picture of the Milky Way. The main features of this view are valid not only for the Milky Way, but for disk galaxies in general.

2.2.1   The Galactic rotation curve

The overall rotation of the Milky Way can only be derived if the movement of objects that closely trace the rotation of the disk can be analyzed throughout the Milky Way. Thus, an extinction-free tracer that is part of the disk (not the halo) is needed. Radioastronomical observations of neutral atomic or molecular gas clouds provide us with such a tracer.

If we assume that a gas cloud moves on a circular orbit around the Galactic Center, the 'Galactic Structure Equation' can be derived:

$$V_{\text{obs}} = R_0 \left(\frac{V_{\text{r}}}{R_{\text{r}}} - \frac{V_0}{R_0} \right) \sin l.$$

Here, V_{obs} is the observed radial velocity, V_{r} is the velocity at distance r, V_0 is the orbital velocity of the Sun, R_{r} denotes the distance from the Galactic Center of a cloud at distance r from the Sun, R_0 is the distance of the Sun from the Galactic Center and l is the Galactic longitude of the cloud at r. We have $V_0 \approx 220 \text{ km s}^{-1}$ and $R_0 \approx 8$ kpc. The radial velocity of the gas cloud can be observed readily and with high precision from spectral lines, e.g. the 21 cm hyperfine structure transition of H I or the 2.6 mm $J = 1 \rightarrow 0$ rotational transition of the CO molecule.

From the Galactic structure equation we expect $V_{\text{obs}} = V_{\text{obs}}(l)$ to be a family of sine curves. If this relation is plotted in a 'longitude–velocity diagram' (e.g. Burton 1895, Hartmann and Burton 1997 for H I, based on the Leiden–Dwingeloo Survey of neutral hydrogen or Dame *et al* 2001 for molecular gas), the expected behavior is indeed seen to a good approximation, with the exception of the region close to the Galactic Center, where the assumption of circular orbits breaks down (see section 2.2.4).

The rotation curve of the Galaxy, i.e. $V_{\text{rot}} = V_{\text{rot}}(R)$, can be derived by considering a special case of the Galactic structure equation, using the classical *tangent point method*. If a cloud is located at the tangent point of the line of

R₀=8.5 kpc
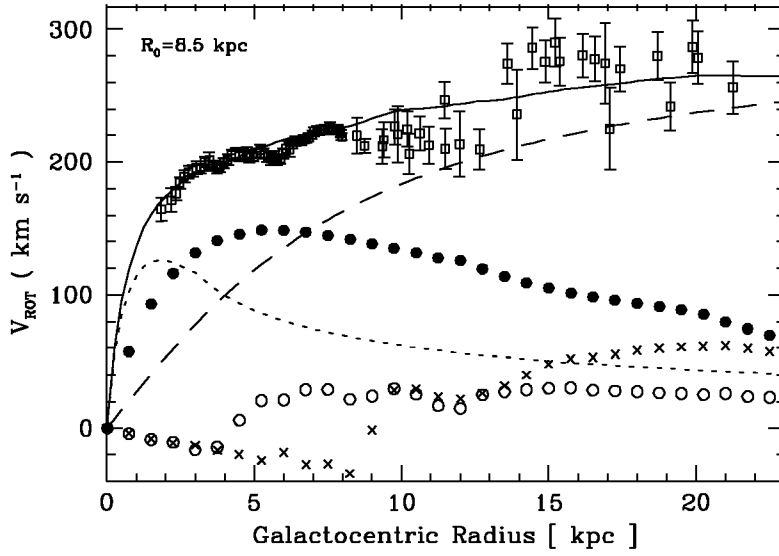
Olling & Merrifield 2000

Figure 2.3. A recent Galactic rotation curve (from Olling and Merrifield 2000). In general, the rotation curve can be described as flat; the slight rise in the outer part is model dependent and changes with, e.g., the assumed distance of the Sun to the Galactic Center. A possible deconvolution into components is given: long dashes, dark matter halo; filled circles, stellar disk; short dashes, bulge; open circles, molecular gas; and crosses, neutral hydrogen.

sight to a (circular) orbit around the Galactic Center, its entire orbital velocity will appear radial. For all realistic mass distributions, this results in the observed 'tangential' velocity being the highest velocity seen in a given spectrum, a quantity that is easy to determine. Then, we have

$$V_{\mathrm{obs,tang}} = V(R_{\mathrm{tang}}) - V(R_0)\sin l.$$

The distance R_{tang} of the tangent point from the Galactic Center is easily calculated as $R_{\mathrm{tang}} = R_0 \sin l$. Thus, the rotation curve can be constructed, at least for $R < R_0$, i.e. within the Solar circle.

In detail, a more careful analysis of non-circular contributions and the used position of the gas clouds is, of course, necessary, since there is not necessarily a gas cloud at any tangent point and non-circular motions, e.g. streaming motions in spiral arms, certainly exist (e.g. Malhotra 1995). The method breaks down in a region within the influence of the bar, i.e. in the inner 2–3 kpc.

Clearly, outside the Solar circle, a tangent point no longer exists. Thus, only V_r/R_r can be determined from the radial velocities. Distances have to be derived

in some other way and are often inaccurate. The difference in error bars between the rotation curve inside and outside the Solar circle is striking (e.g. Olling and Merrifield 2000). Still, all determinations of the rotation of the Milky Way disk agree in one basic fact: after a brief, solid-body-like rise out to $r \sim$ 2–3 kpc, the rotation curve of the Galaxy is *flat* out to $r > 20$ kpc, reaching a maximum value which is obviously close to the Solar rotation velocity, 220 km s^{-1}. Dehnen and Binney (1998) argue that a really flat rotation can only be derived if the distance of the Sun to the Galactic Center is ≤ 7.5 kpc. Otherwise, the rotation curve rises slightly.

The measured rotation curve can be decomposed into contributions due to the various constituents of the Galaxy, most importantly the bulge, the stellar disk, a (neutral and molecular) gas layer and the (dark) halo. Often, a 'maximum disk', i.e. a solution assigning the maximum possible mass to the stellar disk, based on the M/L ratio, is assumed. Even in this case, the only feasible models require a dark halo that dominates the mass budget in the outer part of the Galaxy, roughly outside the Solar circle. A recent example of a possible Galactic rotation curve and its deconvolution is given in figure 2.3.

Flat or even slightly rising rotation curves are a *universal phenomenon* for disk galaxies. Often, they can be more easily determined, based on H I emission, for external galaxies than for the Milky Way. Consequently, a large number has been measured, often to radii well outside the optical disk, since the neutral hydrogen, in many cases, extends beyond the stellar disk (e.g. Bosma 1981, Sofue 1996, 1997). *All* disk rotation curves without exception seem to be flat out to the limits of observability. The only difference between various Hubble types, i.e. spiral galaxies with more or less pronounced bulges, seems to be that 'earlier' Hubble types, Sa and Sb galaxies with stronger bulges, tend to reach a somewhat higher plateau value than later Hubble types (Sc and Sd spirals, Rubin *et al* 1985).

This is the main argument for the universal need for dark matter on galactic scales, a need that seems compelling, as long as one does not want to resort to modifications to the law of gravity (MOND—Modified Newtonian Dynamics— theories, Milgrom 1983 and many subsequent papers) which seem, however, *ad hoc* and unconvincing from a theoretical point of view. Dark matter is also needed, for other reasons, on the larger scale of clusters of galaxies.

A rotation curve is, of course, a global, azimuthally averaged, property of a galaxy. Some hints of the details exist, e.g. in the small 'wiggles' visible in many rotation curves, pointing to streaming motions due to the influence of spiral arms. However, such details dominate the visual appearance of spiral galaxies like the Milky Way.

2.2.2 The disk: spiral arms and their tracers

Theoretically, we can understand spiral arms as compression regions triggered by density waves, which may be either self-excited (Lin and Shu 1964, Lin *et al* 1969) or due to interaction with a companion galaxy, e.g. by swing amplification

of an external perturbation (Toomre 1981). Spiral arms are visually distinctive since they are the sites of the most active star formation. Thus, they host very young stars, belonging to the extreme Population I. In general, metal-rich disk stars are members of Population I, while older, metal-poor halo stars are said to belong to Population II. The existence of a Population III of first-generation stars remains speculative.

Object types certainly belonging to extreme Population I are stars with the earliest spectral types, O and B. These are bright, hot, massive and blue stars which have a lifespan of only a few million years. Therefore, they have no time to disperse, but have to stay close to their birth region within spiral arms. It is their light that lets a spiral arm appear bright and bluish. The youngest open star clusters (age $<$ 10 Myr) also belong to the extreme Population I—bright O and B stars can be part of such clusters or less well-defined 'OB associations'. H II regions, gas clouds ionized by newly born massive and hot stars, are of course also signposts of ongoing star formation, as are the stellar nurseries, the molecular clouds, themselves.

All these objects can, in principle, be used to trace spiral structure observationally. In external galaxies, spiral arms are readily visible and the contrast of dark, filamentary molecular clouds, reddish H II regions and the blue light of massive young stars in close vicinity to each other is particularly impressive. In our own Milky Way the task of locating spiral arms is more challenging. Due to interstellar extinction, explained in section 2.1.2, optical tracers are only useful for local spiral structure, within a distance from the Sun of $\sim$5 kpc. Even within this range, accurate distances are not always easy to determine. Still, an analysis of the distribution of O stars and young open clusters yields several spiral features: the local Orion arm or spur, the inner Sagittarius arm, the outer Perseus arm and a hint of another arm outside the Perseus arm, named Perseus+1 or, simply, 'Outer Arm'. The names of these spiral arms are derived from the constellations towards which they are most clearly seen.

To trace the spiral structure throughout the Galaxy, we need extinction-free tracers. The first such tracer which arrived with the advent of radio astronomy was the 21 cm line of H I. Surveys covering almost the entire Milky Way show a distribution that vaguely resembles a spiral-like structure, but it is not easy to pin down actual spiral arms (e.g. Kerr 1969). One reason for this is, of course, that H I clouds are not extreme Population I objects—they also appear off spiral arms.

Molecular clouds, and especially the 2.6 mm CO line, are better suited to delineating spiral arms, since they combine the advantages of the 21 cm transition—unaffected by extinction, velocity information and, assuming a rotation curve, at least easy-to-obtain kinematic distance information—with their nature as extreme Population I objects highly concentrated in spiral arms. Studies of the large-scale distribution of molecular clouds indeed gave convincing evidence of long, continuous spiral arms (Grabelsky *et al* 1988, Solomon and Rivolo 1989, see figure 2.4). The Sagittarius arm can be shown to connect with an arm in the constellation Carina, extending over an angle of almost 270°. This
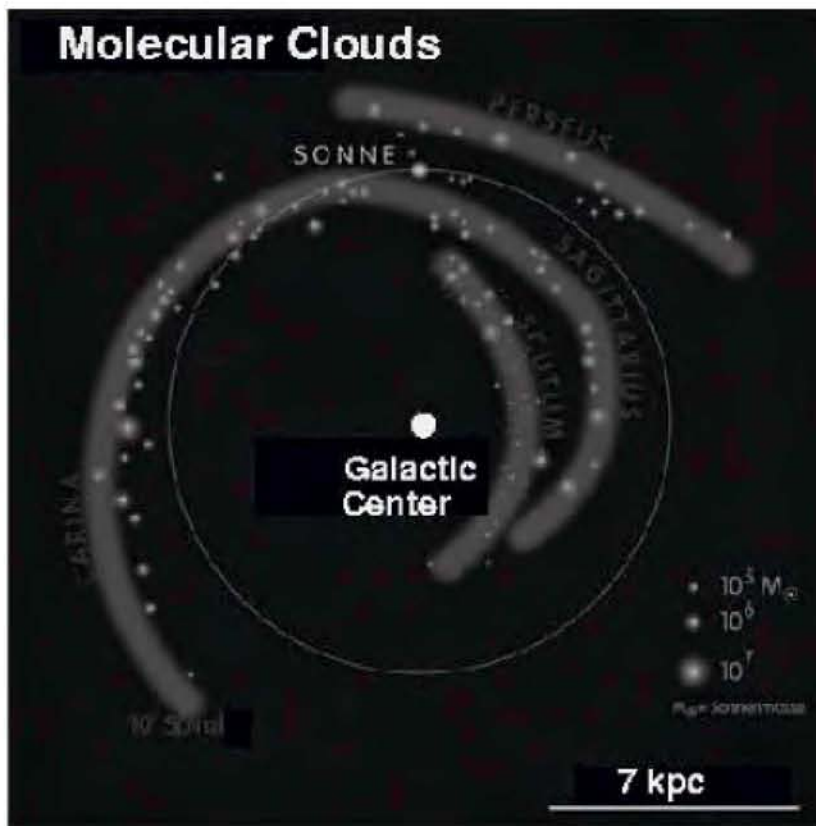
Figure 2.4. The spiral arm structure of the Milky Way derived from the distribution of molecular clouds. Continuous arms become visible. (Image: Hüttemeister/Janssen).

shows that the Milky Way is not a 'flocculent' spiral, an object like M 63 which has a large number of arm fragments which give the overall appearance of a spiral galaxy, but no distinct arms. However, we find too many arm segments for the Milky Way to be (convincingly) classifiable as a 'Grand Design' galaxy like M 51, an object with two very high-contrast spiral arms.

The 'classic' tracer of spiral structure is the distribution of H II regions. This is made possible by the fact that H II regions are not only visible in the optical, but also in radio recombination lines, which are extinction-free tracers allowing the observation of objects on the far side of the Galaxy. A model of the spiral structure of the Milky Way based on the location of H II regions was first constructed by Georgelin and Georgelin in 1976. Data by Downes *et al* (1980) and Caswell and Heynes (1987) were added and collected by, e.g., Taylor and Cordes (1993). In addition to H II region data, tangents to the spiral arms are well defined since

the line-of-sight through the spiral arm is especially long and the density of the relevant tracers very high. Models show the Perseus arm, the Sagittarius–Carina arm, the Scutum–Crux–Centaurus arm and an 'Inner' or 3 kpc arm as continuous features.

Based on such data, most authors envision the Milky Way as a *four-armed spiral* (Vallee 1995). However, more complex models like the superposition of a 2 + 4 arm pattern (Lepine *et al* 2001), also based on analysis of the H II region, i.e. essentially the same data set, supported by stellar kinematics and *N*-particle simulations, are also still discussed. The four-armed nature of the spiral is most certain in young Population I tracers, which tend to be the most luminous objects, which are—as we have seen—those commonly used to define spiral arms. The picture may be different when we examine the distribution of older stars. Drimmel (2000) argues that K-band data, mostly originating from older stars, are well fitted by a two-armed spiral. This may indicate that the Milky Way has a different spiral pattern in the optical and the NIR (or, more physically, in its young and old populations), a phenomenon also seen in a number of external galaxies. We will see in section 2.2.3 how simulations based on the triaxial structure of the bulge also contribute to our understanding of the large-scale spiral structure of the Milky Way, which, however, remains far from complete and perfect.

2.2.3 The bulge: photometric 3D models, bulge/disk models and mass

The bulge in the Milky Way could be seen either as the inner part of the Galactic halo or as the outer part of the Galactic bar, which we will examine in slightly more detail in section 2.2.4. The properties of its stellar population are not easy to determine, since the extinction to this region is generally high. 'Baade's window' is an exception, a region at Galactic latitude $b \approx -4°$, where the extinction happens to be low. Thus, up to a short time ago most observations of the stars in the bulge were made either in Baade's window or other regions at high Galactic latitude ($|b| > 3°$). These observations showed that the bulge in the Milky Way closely resembles other spiral bulges or moderately luminous E or S0 galaxies (e.g. Whitford 1978). Light in the near infrared (NIR) region of the spectrum, at wavelengths of ~ 1.2–2.2 μm, is dominated by old middle and late-type M-giant stars (Blanco *et al* 1984). Thus, the overall photometric and spectroscopic properties of the Galactic bulge are like those of early-type galaxies, dominated by old and metal-poor population II stars, similar to what is found in the Galactic halo.

However, more detailed studies find a gradient in the metallicity (Tiede *et al* 1995), with rising metal content in the direction of the Galactic Center. At the center itself, the metallicity is Solar or higher. Recent NIR observations that came as close as 0.2° to the Center revealed evidence for a bright, young stellar population that can be found only in the inner $\sim 1°$ and quickly declines with increasing radius (Frogel *et al* 1999).

To be able to analyze the structure of the bulge in the Milky Way and draw

Figure 2.5. The bulge of the Milky Way as it appears from the Earth at NIR wavelenths (composite image at 1.25, 2.2 and 3.5 μm spanning 60° in Galactic longitude from the DIRBE instrument aboard the COBE satellite). See also color section.

conclusions about the existence of a triaxial structure, we need to overcome a problem similar to that encountered when trying to derive the spiral structure: if we see the bulge region at all (i.e. at NIR and longer wavelengths), we see it edge-on. Thus, the observed surface brightness distribution has to be deprojected to derive the photometric structure of the bulge. This has been done based on dust-corrected maps from the COBE satellite (DIRBE instrument, in the L (3.5 μm) or K(2.2 μm) band, see figure 2.5). Then, it is possible to model the 3D structure of the bulge and also investigate the consequences the shape of the bulge may have on the non-axisymmetric structure of the disk.

This has been done by, e.g., Binney *et al* (1997), based on a deprojection algorithm developed by Binney and Gerhard (1996). They first obtained a best fit to the observed structure, starting with an analytic model which was improved by using a Richardson–Lucy deconvolution algorithm. The best fit was then deprojected. The result is a clearly and robustly triaxial object with a best-fit axis ratio of 1 : 0.6 : 0.4 and dimensions of 1.8 kpc (major axis) by 1 kpc (minor axis). The angle ϕ between the Sun, the Galactic Center and the long axis of the bulge is also a free parameter of the fit; it turns out to be (again robustly) small: $\phi \sim 20°$. Another parameter of the model is the position of the Sun above or below the plane of the Milky Way, which is 14 ± 4 pc above the plane. Close to the Center, the radial brightness profile is well approximated by a power law, but the 'barred bulge' seems to be embedded in a thin elliptical disk 3.5×2.0 kpc in size. The bar has a pattern speed of $\Omega_P \sim 60$–70 km s^{-1}. At a radius of $\sim$3 kpc, there is a secondary brightness maximum along the y-axis that may be related to the 3 kpc spiral arm.

Similar deprojections and models by other authors (e.g. Freudenreich 1998) have arrived at compatible results.

A slightly different approach was taken by Fux (1997). He used the

COBE/DIRBE K-band map to select the best models out of a large number of stellar dynamical simulations. The models contain a 'nucleus-spheroid' component which represents the bulge, a double-exponential disk and an oblate exponential halo. They extend out to a radius of 10 kpc.

The agreement with Binney *et al*'s results is good: Fux finds $\phi = 28° \pm 7°$, $\Omega_P = 55\pm5$ km s^{-1} and a bar corotation radius of 4.3 ± 0.5 kpc, which constitutes a firm upper limit to the possible bar size. His bar axis ratio is $\sim$0.5.

Examining the families of models, it is found that at a late enough time a bar instability always forms—indeed, in numerical simulations it is very easy to produce persistent bars, some of which are due to numerical effects. Spiral structure is far more rarely seen in these models and, if it appears, it is very short lived. This is partly due to the fact that the models investigated so far are pure stellar dynamical ones, which do not take the influence of a gaseous component into account. Even though most of the mass resides in the stars, gas can have a large effect on the structure that emerges from a model, since it is the coldest, most dissipative component of a galactic system, reacting most readily to perturbations. However, models including gas dynamics are more difficult to realize than purely stellar dynamical calculations.

Still, recently a number of authors have presented bulge/disk models which include gas dynamics (Englmaier and Gerhard 1999, Fux 1999, Weiner and Sellwood 1999). Englmaier and Gerhard calculate quasi-equilibrium flow solutions in the gravitational potential of the deprojected COBE NIR bulge/bar and disk. Their models extend out to a radius of 7 kpc. The best models not only confirm—again—the barred bulge of the earlier deprojections, with $\phi = 20°$–$25°$, $\Omega_P = 55$ km s^{-1} and a corotation radius of 3.5 ± 0.5 kpc; they also quantitatively reproduce the location of the spiral arm tangents determined by a variety of tracers inside the solar circle, spiral arm locations that we have already encountered as the most certain features of the Milky Way spiral structure. The proposed four-armed spiral structure is clearly confirmed and even more pronounced if an additional spiral arm potential is included. A sample model is shown in figure 2.6.

Another assumption mentioned earlier (section 2.2.1) is also supported by gas-dynamical modelling: this is the maximal disk, i.e. a maximal mass-to-light (M/L) ratio, constant for the NIR bulge and disk, that is compatible with the rotation. The implication of this is that the dark matter contribution to the mass in the inner galaxy is small.

We can be fairly certain that the bulge indeed has a triaxial structure since many independent determinations arrive at very similar results. But we have not said anything so far about one very fundamental parameter of the bulge: its mass. 'Conventional' mass determinations based on NIR brightness and a constant M/L ratio yield $M_{bulge} = (1.6$–$2.0) \times 10^{10}\ M_\odot$. This already takes into account the barred, triaxial shape determined earlier.

Interestingly, it is possible that these mass determinations may be contradicted by bulge microlensing experiments (i.e. gravitational lensing by stars, see chapter 9 for a discussion of lensing). In such experiments, millions of
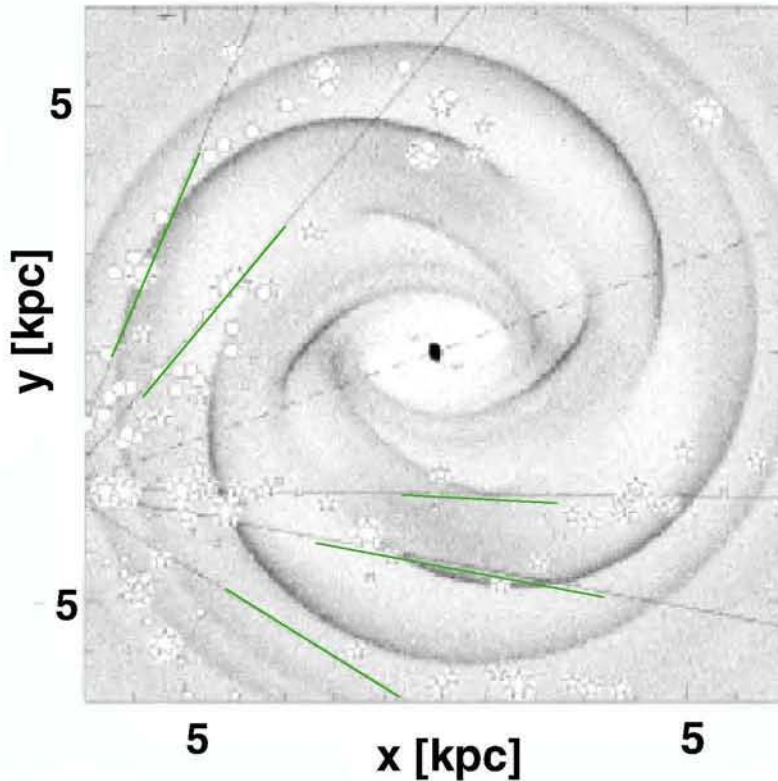
Englmaier & Gerhard 1999

Figure 2.6. A combined gas-dynamical model of the bulge and the disk, including the influence of the dark halo (from Englmaier and Gerhard 1999), calculated with 100 000 particles. The spiral arm tangents are marked; they are well matched by the model which clearly results in a four-armed spiral.

(bulge) stars are monitored for brightness variations. True microlensing events are characterized by symmetric light curves that are independent of color—a secure criterion for distinguishing a microlensing event, when light from a background star is amplified by the close passage of an (unseen) lensing star, from an intrinsically variable star. Microlensing data, which are now based on good statistics and ten years of observations, are collected for a number of purposes, among them the search for Massive Compact Halo Objects (MACHOs) as possible contributions to dark matter and the search for planetary systems. They yield an 'optical depth' toward the bulge of $\tau = (2\text{–}3) \times 10^{-6}$ (Alcock *et al* 2000) for microlensing.

While this is certainly a small value, it is too high to be compatible with bulge

masses determined by the standard method (Binney *et al* 2000, Gyuk 1999). It seems that a bulge mass of $2.5 \times 10^{10} M_\odot$ is required by the microlensing data, which possibly does not agree with any realistic bulge model, though Blum (1995) finds a mass of $2.8 \times 10^{10} M_\odot$ for very special parameters within the framework of a barred stellar distribution. However, in his model the pattern speed required is uncomfortably high; lower values of Ω_P reduce M_{bulge}.

It remains to be seen whether these concerns will result in an upward correction to the bulge mass. Alternatively, it is of course always possible (though not very satisfying) to assume that our line-of-sight toward the bulge is not quite typical.

2.2.4 The nuclear bulge or bar and the Central Molecular Zone

We have now reached the inner 500 pc of the Galaxy, the extended 'Galactic Center' (GC) region, which shows a number of characteristics that are distinctly different from any other part of the Milky Way. Observationally, investigation of the inner 500 pc (corresponding to a Galactic longitude of $l = \pm 3.5°$), at least close to the Galactic plane, is the domain of radio, millimetre and infrared astronomy, due to an extinction that does not allow observations at optical wavelengths. More recently, information on the GC region has been complemented by data from X-ray satellites.

Historically, the first maps of the inner 500 pc were obtained at long cm wavelengths in the radio continuum regime. Even at these long wavelengths, improvements are still possible, as has been impressively shown by the wide-field radio image of the GC obtained at $\lambda = 90$ cm at the Very Large Array (VLA) presented by LaRosa *et al* (2000; see figure 2.7). At these wavelengths, the emission is mostly non-thermal, tracing supernova remnants, some of which are foreground objects, but H II regions and a large number of straight non-thermal filaments are also studied. Many of these filaments are located in the Galactic Center Radio Arc, a projected distance of 50 pc from Sgr A, which hosts the GC itself. The straightness of the filaments strongly suggests a direct relation to the magnetic field structure in the GC region.

The morphology of the ISM in the GC region is dominated by a number of H II regions, the most important of which are historically known as Sgr A, B2, C and D. The very presence of these H II regions shows that star formation in the GC region is ongoing. This conclusion is supported by the results on the age structure of the stellar population in the bulge mentioned earlier and the detection of young and massive star clusters. Not only does the Sgr A region, i.e. the 'actual' GC, host an unusual star cluster, there are at least two more such clusters, the Arches and the Quintuplet clusters (Cotera *et al* 1996, Figer *et al* 1999). These components of the GC region will be discussed in more detail in chapter 4.

In the following, we will focus on the dense, molecular component of the ISM in the GC region. Giant molecular clouds are associated with the H II regions in the GC region. Surveys of the GC region in molecular lines show, however,

Wide-Field Radio Image of the
Galactic Center
$\lambda = 90\ cm$

Sgr D HII

Sgr D SNR

SNR 0.9+0.1

Sgr B2

SNR 0.3+0.0

Sgr B1

Threads

New Feature

Arc

Background Galaxy

Sgr A

Threads

Perpendicular thread?

Sgr C   Coherent
        structure?

Snake

Mouse

Sgr E

SNR 359.0-00.9

SNR 359.1-00.5

~0.5°
~75 pc
~240 light years

Tornado (SNR?)
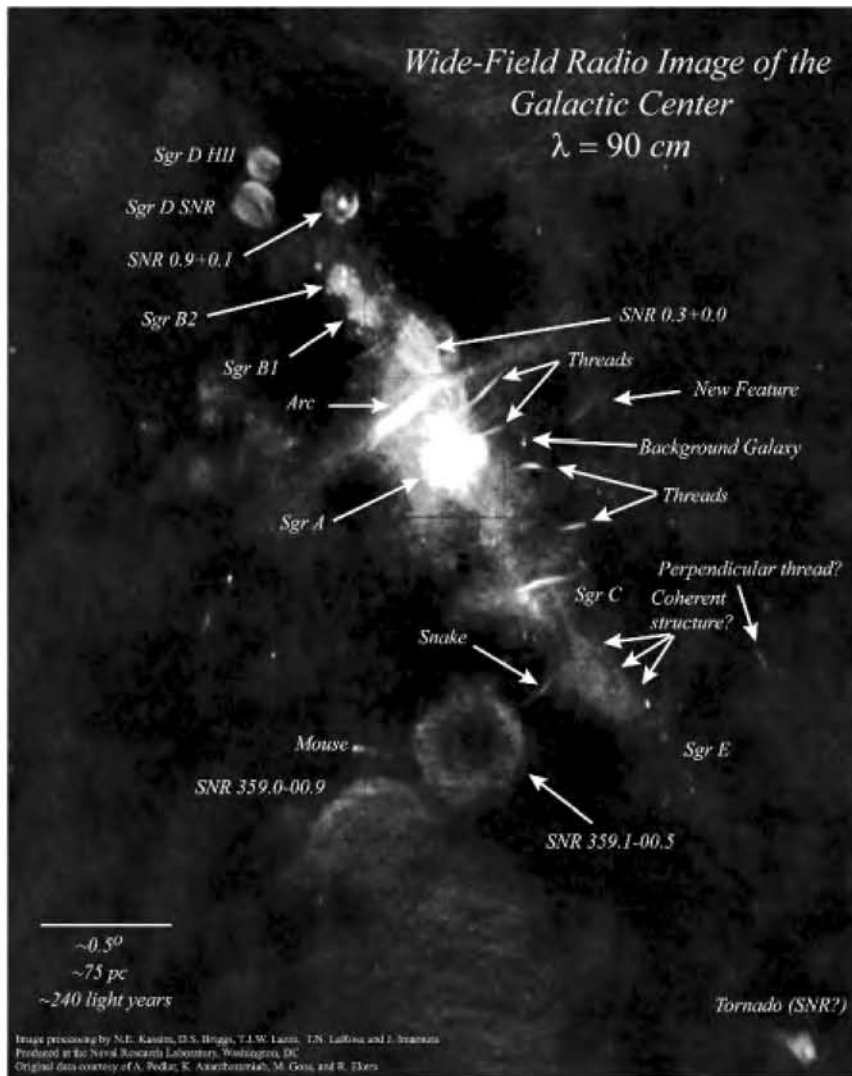
Figure 2.7. Full 330 MHz wide-field image of the Galactic Center at an angular resolution of approximately 45″ taken with the Very Large Array (LaRosa *et al* 2000). See also color section.

that molecular gas is very widespread in the GC region. The gas is traced by spectral lines in the mm range of the spectrum, most importantly the emission of ^{12}CO and its isotopomers (^{12}CO, Dame *et al* 2001; ^{13}CO, Heiligman 1987; $C^{18}O$, Dahmen *et al* 1997). Other useful molecules tracing gas at higher densities ($n(H_2) > 10^4$ cm^{-3}) are, e.g., HCN (Paglione *et al* 1998) or the shock tracer SiO (Martin-Pintado *et al* 1997).

Molecular emission is so widespread that we can speak of a continuous 'Central Molecular Zone' (CMZ) (Morris and Serabyn 1996). Its investigation requires surveys covering several square degrees using (preferentially southern) mm telescopes. These are, even today, large observational projects possibly taking years of observing time, since until very recently radio telescopes could only map a source by observing it point by point, a very time-consuming process for a very extended region like the CMZ. A new generation of array receivers promises significant improvement, even though the number of 'pixels' so far is still typically <20. The large extent of the CMZ therefore explains the somewhat paradoxical situation that we tend to know more about the large-scale molecular content of nearby external galaxies in many molecules and transitions than in the CMZ. On the other hand, any observation of the CMZ, even with small telescopes, has better resolution than what can be achieved even with interferometers for external galaxies. Thus, the CMZ provides us with an important template for the study of external galaxies, and Galactic and extragalactic studies complement each other, a situation that is similar to the study of other constituents of the Milky Way.

The molecular layer of the Milky Way in general, including the CMZ and disk, is the coldest (10 K $\leq$ T_{kin} $\leq$ 100 K), thinnest (scale height <100 pc) and most dissipative component of the Galaxy. Thus, it is most susceptible to gravitational perturbations, caused either by a spiral arm or a bar. A survey of the entire plane of the Milky Way in ^{12}CO (Dame *et al* 2001) shows, as expected, an almost 1 : 1 correspondence between the presence of molecular gas and dark (dust) clouds, i.e. regions of high extinction.

The gas in the CMZ is distinctly different from the molecular clouds found in the disk of the Milky Way: in the innermost Galaxy we find gas that is, on average, warmer and denser than in the disk, with temperatures reaching 100 K and more. In detail, the temperature structure of the CMZ gas is complex—non-LTE (LTE = local thermal equilibrium) radiative transfer models based on multi-line studies show that a large part of the dense component is cool (T_{kin} ~ 20–30 K), while a 'diffuse', warmer component of much lower density is more widespread and may not be bound to molecular clouds (Hüttemeister *et al* 1993, 1998, Dahmen *et al* 1998). Shocks play an important role and are traced by specific molecules like SiO as well as by H_2 emission originating from a hot (for a molecular cloud, T_{kin} ~ 600 K), possibly dense component, which, however, represents only a very small part of the total molecular mass in the CMZ (Rodriguez-Fernandez *et al* 2001).

This mass is not easy to estimate, since it has been shown from line intensity ratio analysis of different CO isotopomers that the Galactic 'Standard Conversion Factor' between ^{12}CO 1 $\rightarrow$ 0 intensity and H_2 column density ($N(H_2) \approx 2 \times 10^{20}$ I_{CO}) systematically overestimates the gas mass if a warm, thin gas component is important. In this case, the ^{12}CO emission is 'overluminous' compared to the Galactic disk conditions for which the conversion factor has been calibrated (Dahmen *et al* 1998). This result also applies to the central regions

of external galaxies, even to Ultraluminous Infrared Galaxies (ULIRGs) which concentrate up to $10^9 M_\odot$ of molecular gas in their inner 500 pc.

Another aspect that makes the investigation of the CMZ (and also its neutral H I counterpart) worthwhile ties in with the determination of the three-dimensional shape of the bulge. We have pointed out in section 2.2.3 that the bulge can be seen as the outer part of the Galactic Center bar. Consequently, we expect to see clear evidence of the bar potential in the molecular and also neutral gas, since this is the most unstable component to gravitational perturbations.

Indeed, very convincing evidence for the barred morphology of the inner Galaxy has been found in studies of the kinematics of the gas. The evidence has accumulated over a decade, starting with work by Binney *et al* (1991). It played a very important role in establishing the Milky Way as a (weakly) barred galaxy, a property it shares with two-thirds of all spiral galaxies. The arguments rest on an examination of the longitude–velocity diagrams of both the CO and H I distribution in the inner 500 pc region. These diagrams show a characteristic 'parallelogram' structure that is not compatible with the gas clouds being on circular orbits around the Galactic Center. Strong emission is found at velocities that are 'forbidden' for circular motion, implying that the gas moves in a non-axisymmetric bar potential.

The parameters of this bar (ϕ and the corotation radius) had already been determined by Binney *et al* to be entirely consistent with what is found for the NIR structure of the bulge. Of course, the molecular bar is flatter than the bulge, since the gas is highly concentrated within the Galactic plane. The more sophisticated gas-dynamical simulations mentioned earlier (e.g. Fux 1999, Weiner and Sellwood 1999, Englmaier and Gerhard 1999) reproduce the detailed structure of the longitude–velocity diagram of the GC region, both for molecular and neutral atomic gas, in impressive detail in models taking into account the bar potential and gas-streaming motions along the bar.

2.2.5 Gas flows and infall: Feeding the nuclear region

A bar provides an obvious mechanism to redistribute angular momentum and to funnel material toward the center of a galaxy. This results in the expectation that gas will flow along the bar and thus become available to feed central activity, a circumnuclear starburst or, possibly, a central black hole. Since the bar is, therefore, intimately connected to the degree of activity we can expect from the Galactic black hole, we will briefly examine the processes that lead to the inflow.

The structure of a bar is governed by the location of its characteristic *resonances*. The most important of those are the corotation resonance (CR, mentioned earlier), where $\Omega_P = \Omega$, i.e. the angular velocity of circular rotation Ω is identical to the bar pattern speed Ω_P, and the Inner Lindblad Resonance(s) (ILR), with the criterion $\Omega_P = \Omega - \kappa/2$ (κ is the epicycle frequency—the epicycle frequency is related to the periodic motion of stars at a given location in the disk compared to their motion relative to a reference frame moving at a constant

circular velocity). While there is always one CR, there may be zero, one or two ILRs, depending on the details of the bar potential and the degree of central mass concentration. Within a bar potential, a number of orbit types can be distinguished (e.g. Binney and Tremaine 1987, Contopoulos and Grosböl 1989). The orbits change the orientation of their major axis by 90° at each resonance.

Gas has to settle on orbits that are stable in the sense of being non-intersecting. Otherwise, the clouds, being viscous, suffer collisions that will result in a quick dissipation of angular momentum and orbital energy, and move them to stable orbits deeper within the potential well of the bar. Thus, for gas clouds we can restrict ourselves to considering two orbit types: x_1-orbits, which are elongated along the bar axis, occur between the ILR and the CR and sustain the bar; and x_2-orbits, which are more circular, but elongated perpendicular to the bar major axis and occur within the ILR. We ignore possible complications due to two ILRs and dynamically detached inner bars. We see at once why the CR limits the bar length: outside the CR the orbits change orientation again and cannot sustain a bar. Also, if x_2-orbits become very dominant, which is the case if the central mass concentration grows very large, the bar self-destructs, since x_2-orbits have an 'anti-bar' orientation.

There is an innermost possible x_1-orbit, with a cusp at its turning point. All x_1-orbits inside of this cusped orbit develop loops and thus are not stable for gas. When passing close to the center, the gas on inner (but still 'allowed') x_1-orbits passes a region close to the ILR where x_2-orbits are already possible. Gas clouds on x_1-orbits may encounter gas that has already settled on x_2-orbits, leading to collisions and a characteristic 'ILR shock'. As a result, gas piles up close to the ILR, leading to either a 'twin peak' (Kenney *et al* 1992) or a ring structure, often with starburst activity.

But not all gas moves from x_1- to x_2-orbits during its first passage close to the ILR. Part of the gas loses some angular momentum in a collision, but not enough to settle on an x_2-orbit. This gas forms a 'spray' of likely unbound molecular gas that moves across the bar and impacts gas still on x_1-orbits on the far side of the bar, causing another region of characteristic bar shocks. Thus, orbital energy is dissipated by shocks, cloud collisions and tidal forces exerted by the bar potential (Das and Jog 1995). All these effects result in a loss of angular momentum and a net inflow of the gas. They also explain the presence of an unbound, diffuse molecular gas component that has been removed from the molecular clouds.

The shape of the characteristic bar shocks, which define the main lanes of infall, can be successfully reproduced in hydrodynamical simulations (e.g. Athanassoula 1992, see figure 2.8). The dust lanes are curved, since due to the viscosity of the molecular ISM the transitions from one orbit type to another are less abrupt than predicted when only considering the bar potential. The shape of the dust lanes agrees very well with observations in strongly barred external galaxies like NGC 1365. Their detailed shape depends on, e.g., the degree of central concentration, the bar pattern speed and the (in)homogeneity of the bar.

Are these concepts applicable to our observations of the CMZ? The ILR
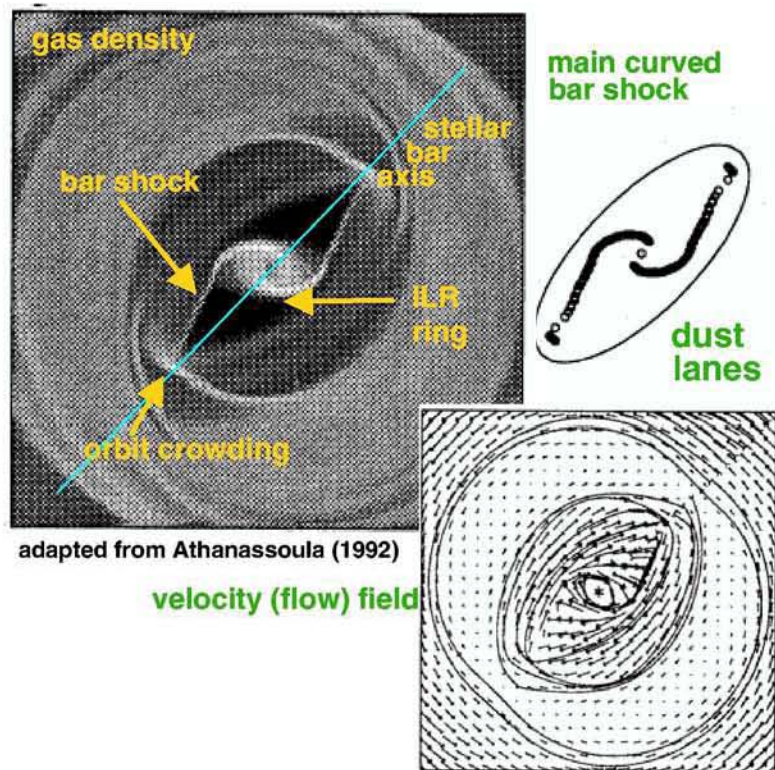
Figure 2.8. Gas flow within a bar (adapted from Athanassoula 1992). The relation between gas density, the location of the characteristic bar shocks and the velocity field are apparent in this example of a hydrodynamical simulation.

region may be identified with the largest pile-up of molecular gas in the CMZ, found in the Sgr B2 region, at a projected distance of ~80 pc from Sgr A. The Sgr B2 region can indeed be considered to be the 'molecular center' of the Milky Way, since it concentrates more molecular mass than the Sgr A complex. It may thus represent part of an (incomplete and very inhomogeneous) ILR ring or one of two twin peaks. It certainly shows evidence of the presence of strong shocks. The Sgr B2 region is located north of the GC. The Sgr C complex may be its southern counterpart, but this Giant Molecular Cloud is considerably less massive. In general, the distribution of molecular gas in the CMZ is significantly asymmetric, with more material found at northern longitudes. This points to a (probably transient) $m = 1$ instability in addition to the well-known bisymmetric $m = 2$ (bar) instability.

Is it possible to trace the dissipation of energy and the shock activity along the bar directly in the molecular component? The kinematics of the gas can indeed be linked to the 'chemistry' by observing characteristic shock tracers like SiO. This molecule is only encountered in the gas phase if it has been removed from dust grain surfaces by shocks. Thus, regions with enhanced SiO abundance can be identified with shocks. SiO emission in the Galactic disk is highly confined to stellar outflows, while in the CMZ the SiO emission can be mapped on large scales. A detailed investigation (Hüttemeister *et al* 1998) shows a clearly enhanced abundance close to the ILR (Sgr B2) and along the regions where the impact of the 'spray' on the gas remaining on x_1-orbits would be expected.

Thus, a gas flow along the bar at least to the ILR region, identified with Sgr B2, within the concept of gas moving on classical bar orbits outlined earlier, is convincingly observed. Further transport inward, necessary to feed the black hole, may be less easy, since the efficiency of energy dissipation will not be as high as in the sphere of influence of the rounder, non-intersecting x_2-orbits. Still, further infall certainly takes place. One process that has to operate is dynamical friction. It can be estimated that gas is fed to the very center of the Galaxy on timescales that are far shorter than a Hubble time (timescale of cosmological evolution or 'age' of the universe). However, the inflow to the ILR and possible activity cycles related to the black hole do not have to be coupled.

The fate of the gas closer to the black hole will be followed in chapters 4 and 10.

2.3 Galaxy evolution

Having seen what constitutes our Milky Way, i.e. 'what is there', we will now briefly explore the question how 'it got there', i.e. how a galaxy forms and evolves. Clearly, the field of galaxy formation and evolution is huge, and far beyond the scope of this chapter. Thus, we will only sketch the barest, qualitative outlines of the most crucial concepts and results.

The study of galaxy evolution has gained enormous momentum since very deep observations of almost randomly selected, small areas of the sky at a number of wavelengths have become available. The most famous of these 'Deep Fields' are the Hubble Deep Fields (HDF North, observed in 1995, and HDF South, observed in 1998, see Ferguson *et al* 2000 for a review of the HDF observations and their impact). Deep fields, often centered on the HDFs, have now been obtained in a multitude of wavelength ranges, i.e. in the far infrared (ISO, e.g. Serjeant *et al* 1997), in the submillimetre (Scuba, e.g. Hughes *et al* 1998) and at X-ray wavelengths (Chandra: Giaconni *et al* 2001).

In principle, these images allow the observation of galaxy evolution 'as it happens', since they contain both very distant and more nearby objects. Thus, it is possible to follow the evolution of the galaxy population. Of course, to do so, it is necessary to distinguish between true evolutionary effects and morphological

variations. Thus, a measure of distance, or redshift, is needed. It is difficult or impossible to determine spectroscopic redshifts for very faint objects, but the problem has been moderated with the advent of much more easily measured photometric redshifts. It could be established that these redshifts are indeed fairly accurate (e.g. Fernandez-Soto 2001).

2.3.1 Hierarchical, bottom-up structure formation

The investigation of large-scale structure formation is the domain of numerical simulations of dark matter particles, using some of the largest parallel supercomputers in existence. For example, the 'Hubble Volume' project of the international Virgo collaboration (see www.mpa-garching.mpg.de/~virgo/virgo) follows 10^9 dark matter particles from $z \sim 5$ ($z \sim 1.4$ for a somewhat smaller volume) to the present. Thus, the development of galaxy clustering (Evrard *et al* 2001, Colberg *et al* 2000) or the mass function of dark matter halos (Jenkins *et al* 2000) can be investigated.

The simulations reproduce observational redshift surveys such as the 2dF survey (Percival *et al* 2001), the Las Campanas redshift survey (e.g. Shectman *et al* 1996) and—soon—the Sloan Digital Sky Survey to a reasonable degree of accuracy. At high redshifts, they show little large-scale structure, though of course the seeds of structure can be tied in with the Cosmic Microwave Background fluctuation spectrum, and beyond, as established by the COBE satellite and the recent Boomerang (e.g. de Bernardis *et al* 2000) and Maxima (Hanany *et al* 2000) balloon experiments.

The development of filamentary structures ('bubbles and voids') is striking and can already be followed in simulations with somewhat fewer particles, like the one by Jenkins *et al* (1998), based on 1.7×10^6 particles, which was carried out assuming four different cosmological models (see figure 2.9). Universes with low matter content ($\Omega_M = 0.3$) such as the currently favoured ΛCDM cosmology form structure earlier than universes with $\Omega_M = 1$, like the formerly 'standard' Einstein–de Sitter universe, that seems to be ruled out by current data. At low redshift, all simulations give similar results. They probe size scales down to ~ 10 kpc. Thus, they include the strong clustering regime and follow the hierarchical formation of clusters, but do not resolve the evolution of individual galaxies. The simulations have predictive power and help decide between cosmological models; for example an analysis of the Hubble volume leads to the prediction that too many hot (in the sense of X-ray temperatures) clusters like the Coma cluster at $z \sim 1$ would not be compatible with a ΛCDM-based model.

On the slightly smaller scale of the evolution of galaxy clusters, we expect the cluster environment itself to influence the evolution of the member galaxies. Smaller building blocks are expected to merge into larger objects, and this process can again be followed in simulations (e.g. Moore *et al* 1999). It turns out that, while hierarchical merging certainly occurs, the extent of the merging depends
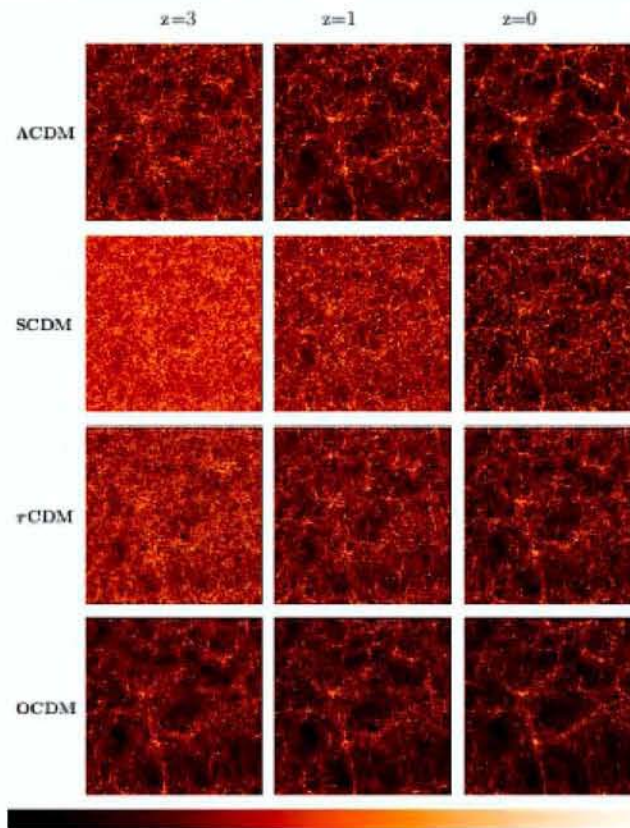
The VIRGO Collaboration 1996

Figure 2.9. Four simulations of large-scale structure formation by the VIRGO consortium (Jenkin *et al* 1998). Four different cosmological scenarios are assumed; the development of structure is depicted at redshifts $z = 3$, 1, and 0. The boxsize is 239.5 Mpc/h.

on the velocity profile and dispersion of the cluster. In any case, small dark matter subhalos seem to survive in surprisingly large numbers down to the scale of individual giant galaxies. In simulations, the halos of individual galaxies look very similar to galaxy clusters, and this result persists at the highest currently possible resolution of numerical, dissipationless N-body simulations (Ghigna *et al* 2000).

Thus, hierarchical models for structure formation naturally form massive dark matter halos with a wealth of substructure. On a cluster scale, this substructure can easily be identified with visible galaxies, that are (with some

bias function that is difficult to determine) hosted by the dark matter subhalos. The observed distribution of galaxies in clusters agrees reasonably well with prediction.

On the scale of individual galaxies, however, there may be a problem: here, the subhalos correspond to satellite dwarf galaxies, and the observed abundance of such dwarf satellites does not match the large number of predicted dark matter subhalos (Moore *et al* 1999, Klypin *et al* 1999). It remains to be seen whether this is a serious problem of the otherwise very successful Cold Dark Matter (CDM) simulations or whether it can be solved by fine-tuning the models, as suggested by, e.g., Font *et al* (2001), who argue that at least the dynamical impact of the subhalos on the thin disk of a galaxy should be minor.

Evidence for hierarchical, bottom-up galaxy formation originates not only from numerical simulations; it is strongly supported from an observational viewpoint. Analysis of deep field data shows that the morphological classification of galaxies by the classical Hubble sequence breaks down at redshifts >1 (e.g. Driver *et al* 1998). Barred galaxies seem to become rare at even lower redshift (Abraham *et al* 1999) while the number of 'irregular' or 'peculiar' galaxies increases steeply (see the recent review by Abraham and van den Bergh 2001).

In general, the size–redshift relation seen for E/S0 and spiral systems seems to point to their assembly of redshifts >1. There also seems to be an excess of faint blue, very compact galaxies, many of which are located at $z \sim 0.5$, but some are at $z > 2$. This appears to be an actively evolving galaxy population, constituting, at least in part, the building blocks of larger systems.

In some cases, these subsystems or building blocks may have been caught in the act of taking part in a hierarchical merging process. Pascarelle *et al* (1996) found 18 small, bluish objects in a ~ 0.7 Mpc field at $z = 2.39$ and these were interpreted as the building blocks of a future large galaxy. Similarly, Campos *et al* (1999) reported the detection of 56 Lyman α emitters in a small field adjacent to a quasi-stellar object.

2.3.2   Evolutionary mechanisms: mergers and 'internal' processes

Most galaxies are not isolated—as we have seen, they tend to form in clusters, or at least in groups. This cluster environment is expected to influence not only the number of galaxies directly through merging, but also their type. This effect, dubbed 'galaxy harassment', probably has consequences for the balance of Hubble types in a cluster between redshifts of, e.g., $z \sim 0.4$ and the present, that is, in a redshift regime where the Hubble sequence still describes the galaxy type adequately. It seems that more distant clusters have a larger relative fraction of small spiral galaxies, many of which show some indication of high star formation or starburst activity, than clusters in the local universe. The latter are dominated by spheroidal galaxy types, ellipticals and S0s (Moore *et al* 1998).

Merging itself is, of course, the most dramatic and obvious driving force of galaxy evolution. In a scenario of hierarchical structure formation, the merger

rate is expected to increase with redshift proportional to $(1 + z)^m$ with $m = 2$–3. This relation is expected to hold at least up to $z = 2$–3.

There are many excellent simulations of galaxy mergers (e.g. Barnes and Hernquist 1996, Mihos and Hernquist 1996, Springel *et al* 2001). It is important for the outcome to take into account the role of gas and (if possible) the feedback of the star formation process (see Kauffmann *et al* (1999) for a prescription). The simulation of collisionless dark matter or stellar particles can only be a first step in such simulations. The morphology of interacting and merging galaxies is well reproduced by models, down to the tidal tail structure of individual real-life objects. Usually, the models suggest that the merger remnant looks much like an elliptical galaxy. Specifically, the remnant structure follows the well-known $R^{1/4}$ law for the surface brightness of an elliptical fairly well, though it somewhat depends on the initial conditions and there may be discrepancies in the details.

This has led to the conclusion that ellipticals are the endpoints of galaxy evolution through mergers. During the merger, the galaxies pass through a phase of a very intense central starburst, since the gas is concentrated quickly into the nuclear region. Briefly, they may shine brightly as ULIRGs (ultra-luminous infrared galaxies), emitting more than $10^{12} L_\odot$ in the Far Infrared. During this phase, most of the gas of the progenitor galaxies is consumed. Observations have indeed shown that all ULIRGs seem to be mergers, often even multiple ones. Once the starburst is over, ellipticals (and spiral bulges) evolve only passively, i.e. by the ageing of the stellar population. If they have largely assembled in a 'merger age' at $z \sim 2$, this explains naturally why most ellipticals and bulges today appear to be old, reddish objects.

Apart from the spectacular evolution by merging, 'internal' mechanisms may also lead to secular changes in the appearance of galaxies. These are slower evolutionary mechanisms that take place after the initial assembly and most easily work on galaxies which have not become ellipticals, but are instead gas-rich disk galaxies. Interactions which do not lead to mergers play an important role in triggering or at least speeding up these evolutionary processes.

In contrast to the passively evolving ellipticals, the disk galaxies continue to form stars at a fairly constant rate of a few $M_\odot \mathrm{yr}^{-1}$. This star formation takes place, as we have seen, mostly in spiral arms, and possibly in the central region, especially if it is fed by a bar. The detailed structure of the spiral arms themselves is almost certainly subject to secular changes, even in Grand Design spirals, though the general character of a disk galaxy as a spiral remains unchanged. What is the gas supply for continuing star formation? Some of the necessary replenishment takes place by inflow from the more gas-rich outer regions of the disk—we have seen that the H I disk often extends far beyond the optical disk. It is, however, also possible that infall of intergalactic (intra-cluster or intra-group) or halo gas clouds takes place. Chemical evolution models and the star formation and thus the gas consumption rates in many disk galaxies may require some infall.

Evolution along the Hubble sequence may happen to some degree. If so, then the direction of evolution is late $\rightarrow$ early, since all processes result in a higher

Shaver et al. 1999   **Cattaneo 2001**
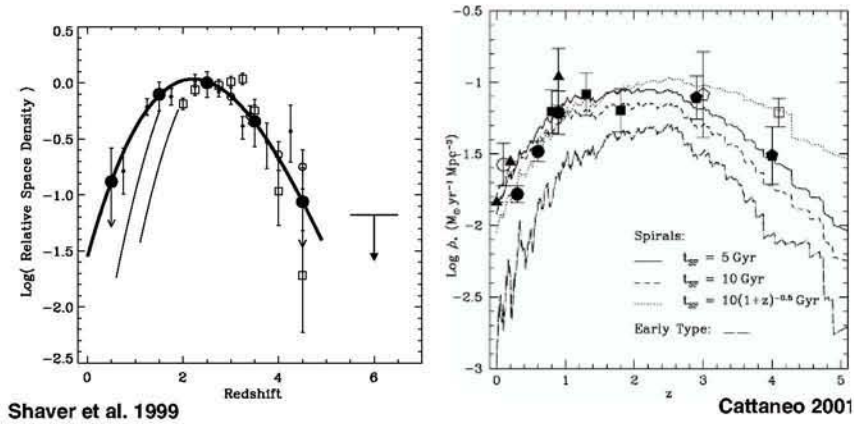
Figure 2.10. Left panel: the space density of a quasar has a maximum at a redshift of ~ 2 (from Shaver *et al* 1999). Right panel: the same may be true for the star formation rate, but in this case, the decline beyond $z \sim 2$ is not well established (Cattaneo 2001; the points with error bars correspond to measurements; the curves are predictions for different evolutionary models).

central mass concentration and thus a more pronounced bulge. A bar, whether transient or persistent, should usually be involved in such a process, since it is the most efficient means of angular momentum transport.

More dramatic events, i.e. major mergers between disk galaxies with a resulting strong starburst, may of course also take place in the local universe, and we know a number of examples (e.g. Arp 220 and similar objects). Major mergers are, however, rare in the present-day cosmos. Accretion of smaller galaxies is a far more frequent process. In such a 'minor' merger, a large disk (or elliptical) galaxy swallows a smaller companion. It is likely that the Milky Way has been involved in several such acts of cannibalism during its history. At present, it is performing another one: it is in the process of consuming a small dwarf galaxy, the Sgr dwarf, which has already been disrupted and stretched out to a degree that made its very detection difficult (Ibata *et al* 1994).

The growth and evolution of a spiral bulge, e.g. through infall or inflow, may also have consequences for the central black hole, possibly by regulating its rate of mass accretion.

We have seen that the merging rate was certainly higher in the past, and may have had a maximum around $z \sim 2$. Possibly related to this, there is undisputed evidence for a 'Quasar Epoch' at the same redshift (Shaver *et al* 1996, 1999). Quasars are thought to be powered by the most luminous supermassive black holes in the universe. Not only was the true quasar space density at $z \sim 2$ more than an order of magnitude higher than it is now, it is also clear from, e.g.,

complete samples of radio-loud quasars that it declined rapidly at redshifts > 2.5 (see figure 2.10, left-hand panel).

Investigations of the star formation history of the universe, pioneered by Madau *et al* (1996), also show a clear rise by more than a factor of 10 from the present to $z \sim 1$. It is less clear, however, whether the star formation rate declines at $z > 2$–3 (e.g. Cattaneo 2001, see figure 2.10, right-hand panel). This depends, among other factors, on the role of dust extinction in the high-z starburst galaxies and how a population of extremely dusty starbursts that appears in sub-mm-wave deep fields, and seems to have enormous star formation rates, is taken into account.

In any case, the rise in the quasar space density, the merger rate and the star formation rate almost 'in lockstep' is very suggestive of a scenario where the quasar brightness is explained by high accretion rates on massive central black holes, which were assembled at roughly the same time in frequent mergers of galaxy bulges, going along with massive starbursts. Consequently, the black holes experienced the bulk of their growth in this period, and might have grown only slightly since their initial assembly. This is indicated by the strong decrease in quasar light originating in the local universe.

2.4 The relation between black holes and bulges

The evolutionary findings outlined in the previous section suggest the possibility of a relationship between black hole characteristics and bulge parameters. Evidence for this concept and some of its implications will be briefly discussed in the following final section of this introductory chapter.

2.4.1 Black hole mass and bulge mass/luminosity

In recent years, there has been mounting evidence that black holes may be a normal, possibly ubiquitous, component of galaxies: if a galaxy has a bulge or is an elliptical, we expect to find a massive black hole in its center.

The masses of the suspected black holes have been obtained (or estimated) by a variety of methods, some more reliable than others. The mass determination for the Galactic black hole is, of course, outstanding in its accuracy and resolution, and is discussed in detail elsewhere in this book (chapter 8). Very reliable black hole masses can also be obtained by very high resolution radio-interferometric observations of maser disks around central massive objects. Unfortunately, this method can be applied to only a very few objects, since a special disk geometry is required. In the best case, the Keplerian disk rotation under the influence of the central point source can be measured directly, making the estimate for the black hole mass, $M_\bullet$, very secure. This has, however, only been achieved in one case, NGC 4258 (Miyoshi *et al* 1995, Herrnstein *et al* 1999, Bragg *et al* 2000).

The determination of the velocity structure of a central dust disk yields another reliable mass estimate. Of course, the galaxy under investigation must

have such a very central dust disk that can be studied without too much extinction. This restricts the usefulness of the method to a few elliptical galaxies with central disks of dust and ionized gas that can be resolved (at least by the HST, the Hubble Space Telescope). The black hole masses determined for M 84 (Bower *et al* 1998) and NGC 7052 (van der Marel and van den Bosch 1998) are good examples of a successful application of this method.

Most black hole masses have been estimated by means of stellar dynamics, from either HST or ground-based data. Here, the goal is to measure the rise of the velocity dispersion due to the influence of the central mass. In principle, this method can be—and has been—applied to many galaxies, and the first study of black hole demographics was based on values for $M_\bullet$ determined by stellar dynamics (Magorrian *et al* 1998). However, sufficient resolution is an issue here—the 'sphere of influence' of the black hole has to be clearly resolved. This is a difficult task: in a number of cases the rise of the velocity dispersion assigned to the black hole was only observed in the central pixel.

A relatively new and very promising technique for determining $M_\bullet$ is reverberation mapping of Active Galactic Nuclei (AGNs). In this method, time delays between brightness variations in the continuum and in the broad line emission region (BLR) are interpreted as the light travel time between the black hole and the BLR, i.e. the black hole accretion disk. Velocity information comes from the width of the emission lines, and thus the black hole mass can be determined as $M_\bullet \approx V^2 r / G$. The obvious advantage of this method is that it is independent of distance. Thus, obtaining $M_\bullet$ for a large number of AGNs, mostly too far away for the more classical techniques, becomes possible (see Gebhardt *et al* 2000a, Kaspi *et al* 2000, Wandel *et al* 1999). It now seems that this method yields results that are very consistent with reliable determinations by other means.

In any case, in recent years, enough black hole masses have been determined for it to become worthwhile to look for relations between $M_\bullet$ and host galaxy parameters, especially those of the oldest component of a galaxy, the bulge.

The beginnings of what is now known as the relation between bulge and $M_\bullet$ were rather humble and go back to at least Kormendy (1993). There seemed to be evidence that $M_\bullet$ scales with the luminosity (equivalent to the mass for a constant M/L ratio) of the bulge. In a 1995 review paper, Kormendy and Richstone displayed a diagram relating the mass of the bulge to $M_\bullet$ for eight galaxies (and two non-detections). Excluding the mass determination for the Galactic black hole, which seemed to have a mass that was too low in relation to the other estimates, they found a proportionality of $M_\bullet / M_{\mathrm{bulge}} \sim 3 \times 10^{-3}$. The Galactic black hole alone yielded $M_{\bullet,\mathrm{gal}} / M_{\mathrm{bulge,gal}} \sim 1.7 \times 10^{-4}$. Even then, the question could be asked why the best-determined value (for the Galaxy) was excluded from the relation. But at that time, the relation was based on very few galaxies and had to be considered more as a trend than a firm proportionality.

In their paper on the demography of black holes in galaxy centers, Magorrian *et al* (1998) investigated the stellar kinematics of 32 galaxies. After careful modelling of the bulge kinematics to obtain $M_\bullet$, they found a relation between

$M_\bullet$ and M_{bulge} of the form

$$\log M_\bullet = (-1.79 \pm 1.35) + (0.96 \pm 1.12) \log M_{\text{bulge}}.$$

From individual galaxy data, a proportionality of $\langle M_\bullet/M_{\text{bulge}}\rangle \sim 0.016$ can be derived if the unrealistically high value found for NGC 4486B is included. If this galaxy is excluded, the ratio drops to 7×10^{-3}.

Following these results that seemed to indicate that a relation between black hole and bulge mass really existed, albeit with a substantial scatter, a potential problem was pointed out by a number of authors (e.g. Wandel 1999, Ho 1999). The new masses determined by reverberation mapping seemed to be systematically low compared to the stellar dynamical masses. In addition, the large ratio suggested was not compatible with the black hole mass density inferred if the optical QSO luminosity was to be reproduced by (standard) accretion onto black holes.

2.4.2 Black hole mass and bulge velocity dispersion

A revision of the $M_\bullet$–M_{bulge} relation and a possible breakthrough became public in two letters published back to back in the same issue of *Astrophys. J.* Ferrarese and Merritt (2000) and Gebhardt *et al* (2000b) independently presented a new, much tighter relation between $M_\bullet$ and the bulge velocity dispersion σ (see figure 2.11).

Ferrarese and Merritt first sorted black hole masses from the literature into groups of 'reliable' and 'unreliable' values, with many of the (especially ground-based) masses from the Magorrian *et al* sample in the latter category, since Ferrarese and Merritt suspected that in many cases the black hole's sphere of influence was unresolved. However, the scatter in the $M_\bullet$–M_{bulge} relation did not decrease when only the better determined masses were considered. This changed dramatically when σ instead of the bulge mass was correlated with $M_\bullet$. While the scatter was still large for the galaxies with the uncertain black hole masses, the relation for the 12 galaxies with 'good' values for $M_\bullet$ became very tight. Gebhardt *et al* (a large group of scientists overlapping strongly with the authors of the Magorrian *et al* work) found exactly the same effect for a different sample of galaxies and values for $M_\bullet$ based on stellar dynamics, mostly done with very high resolution data obtained from the HST.

The relationship is so tight that it is statistically 'perfect' in the sense of being compatible with only having errors due to measurements, i.e. no intrinsic scatter. This is almost 'embarrassingly' tight (Merritt and Ferrarese 2001c), since a relation this close has to be explained with effects of galaxy assembly and evolution. If confirmed, this relation may be used to predict black hole masses accurately from the much more easily obtained bulge velocity dispersion. The black hole masses found are significantly, by up to an order of magnitude, lower than the masses claimed before. Thus, the ratio between $M_\bullet$ and M_{bulge} (which of course still has a larger scatter) has gone down to a value $\sim 2 \times 10^{-3}$. This solves
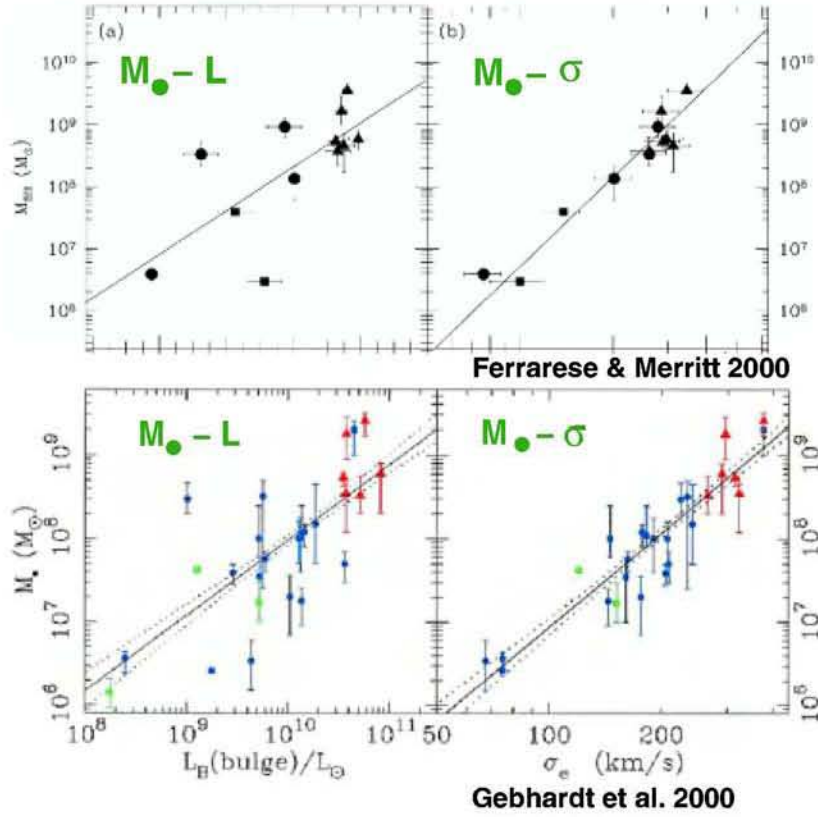
Figure 2.11. Upper panels: the $M_\bullet$–σ relation according to Ferrarese and Merritt (2000). The left panel shows the large scatter in the $M_\bullet$–L_{bulge} relation. For the same sample of galaxies, the scatter is greatly reduced if $M_\bullet$ is related to σ instead. Lower panels: Gebhardt *et al* (2000b) come to the same conclusion, using a different sample of galaxies.

a number of problems: first, the extremely well-determined mass of the Galactic black hole is now almost consistent with the general relation. Second, and maybe more fundamentally important, the density of black holes in the local universe has been redetermined to be $\rho_\bullet \approx 5 \times 10^5 M_\odot$ Mpc^{-3}. This is to be compared to the prediction of the QSO luminosity function: $\rho_\bullet \approx (2\text{–}3) \times 10^5 M_\odot$ Mpc^{-3}. Thus, very little, if any, 'invisible' black hole growth since the quasar epoch is required.

Despite the general agreement, some problems remain. The relations found by the two groups differ in detail: Merritt and Ferrarese find

$$M_\bullet = 1.4 \times 10^8 \, M_\odot \left(\frac{\sigma}{200 \text{ km s}^{-1}} \right)^{4.8 \pm 0.5}$$

while Gebhardt *et al* obtain

$$M_\bullet = 1.2 \times 10^8 \, M_\odot \left(\frac{\sigma}{200 \text{ km s}^{-1}} \right)^{3.8 \pm 0.3}.$$

The different slope may be partly due to a different sample. However, if the relation really has no (or extremely low) intrinsic scatter, it should of course be identical no matter what sample is chosen. There are also differences in the statistical analysis of the data (Merritt and Ferrerese 2001a, b), which contribute to the discrepancy.

It is, at this time, difficult to say whether these differences (in data, method and opinion) only concern details which will be easily sorted out when more and better data become available or whether they touch on fundamental difficulties. Part of the current discussion concerns a possible low-mass cut-off of the relation, with the test case of the Local Group galaxy M 33, a galaxy that has (almost?) no bulge and is sufficiently nearby to allow very sensitive searches for the signature of a black hole. No black hole has been found so far down to the very low mass limit of $M_\bullet < 3000 \, M_\odot$ (Merritt *et al* 2001). Obviously, the implications of such a mass cut-off for the $M_\bullet$–σ relation are related to its slope.

In any case, the impact of the $M_\bullet$–σ relation on studies of galaxy formation and early evolution will very likely be significant. Clearly, it 'must be telling us something fundamental about origins (galaxy formation) and the connection between black hole mass and bulge properties' (Merritt and Ferrarese 2001a). If the relation is simply interpreted as the $M_\bullet$–M_{bulge} relation recast in another form, it implies the well-known Faber–Jackson law for elliptical galaxies (which relates the velocity of stars to the bulge luminosity), and thus corresponds to a (not entirely explained) fundamental plane[1] relation, even with approximately the right slope ($M_{\text{bulge}} \sim \sigma^5$).

However, as we have seen, the $M_\bullet$–σ relation appears much tighter. As a constraint, it has to survive subsequent mergers in the scenario of hierarchical structure formation without being disrupted. Fundamental plane relations also survive merging, but are generally less tight. To add to the 'problem', it appears that another very close relation may have been found very recently: Graham *et al* (2001) report that the correlation between $M_\bullet$ and the light concentration within a bulge's half-light radius is as tight (in fact, statistically slightly tighter) as the $M_\bullet$–σ relation. Since the central condensation can be determined from surface photometry alone, it might be possibly to estimate $M_\bullet$ from fairly simple observations.

In any case, it is probably necessary to look for an additional feedback mechanism during bulge formation that directly connects the black hole mass, the stellar velocity dispersion and possibly (if the Graham *et al* result is confirmed) the central light concentration.

[1] Fundamental plane: the parameter space spanned by the luminosity density, radius, and average kinetic energy of stars in an elliptical galaxy is a plane.

Some possibilities for explaining the $M_\bullet$–σ relation have been suggested, one even before its discovery (Silk and Rees 1998). They propose that black holes form even before bulges from the collapse of $\sim 10^6 M_\odot$ gas clouds, which then accrete and radiate. This drives a wind which acts back on the accretion flow. In this scenario, the black holes have to form very early. Alternatively, black holes can be thought to grow naturally in mergers. The feedback may then be supplied by gas cooling balanced by energy input through supernovae. The impact of this energy input may be stronger for smaller galaxies with shallower gravitational potentials (Kauffmann and Haehnelt 2000, Haehnelt and Kauffmann 2000). Burkert and Silk (2001) also consider black hole growths in mergers, but suggest a somewhat different feedback mechanism: in their model, the accretion is halted when star formation begins to exhaust the gas supply in the outer accretion disk.

All these models result in an $M_\bullet$–σ relation with roughly the expected slope, at least after some fine-tuning. However, they all have problems in explaining why the relation appears so tight. Thus, much work, both observationally and theoretically, remains to be done until the role of the $M_\bullet$–σ relation in the process of galaxy evolution is fully understood.

References

Abraham R G *et al* 1999 *Mon. Not. R. Astron. Soc.* **309** 208
Abraham R G and van den Bergh S 2001 *Science* **293** 1273
Alcock C *et al* 2000 *Astrophys. J.* **541** 734
Athanassoula E 1992 *Mon. Not. R. Astron. Soc.* **259** 328
Barnes J E and Hernquist L 1996 *Astrophys. J.* **471** 115
Binney J, Bissantz N and Gerhard O 2000 *Astrophys. J.* **537** L99
Binney J and Gerhard O 1996 *Mon. Not. R. Astron. Soc.* **279** 1005
Binney J, Gerhard O and Spergel D 1997 *Mon. Not. R. Astron. Soc.* **288** 365
Binney J, Gerhard O, Stark A A, Bally J and Uchida K 1991 *Mon. Not. R. Astron. Soc.* **252** 210
Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press)
Blanco V M, McCarthy M F and Banco B M 1984 *Astrophys. J.* **89** 636
Blum R D 1995 *Astrophys. J.* **444** L89
Bosma A 1981 *Astrophys. J.* **86** 1825
Bosma A, Byun J, Freeman K C and Atthanassoula E 1992 *Astrophys. J.* **400** L21
Bower G A *et al* 1998 *Astrophys. J.* **492** L111
Bragg A, Greenhill L J, Moran J M and Henkel C 2000 *Astrophys. J.* **535** 73
Burkert A and Silk J 2001 *Astrophys. J.* **554** 151
Burkert A and Smith G H 1997 *Astrophys. J.* **474** L15
Burton W B 1985 *Astron. Astrophys. Suppl.* **62** 365
Campos A *et al* 1999 *Astrophys. J.* **511** L1
Caswell J L and Heynes R F 1987 *Astron. Astrophys.* **171** 261
Cattaneo A 2001 *Mon. Not. R. Astron. Soc.* **324** 128

Colberg J M *et al* 2000 *Mon. Not. R. Astron. Soc.* **319** 209

Contopoulos G and Grosböl P 1989 *Astron. Astrophys. Rev.* **1** 261

Cotera A S *et al* 1996 *Astrophys. J.* **461** 750

Crezé E, Chereul E, Bienayme O and Pichon C 1998 *Astron. Astrophys.* **329** 920

Dahmen G *et al* 1997 *Astron. Astrophys. Suppl.* **126** 197

Dahmen G, Hüttemeister S, Wilson T L and Mauersberger R 1998 *Astron. Astrophys.* **331** 959

Dame T M, Hartmann D and Thaddeus P 2001 *Astrophys. J.* **547** 792

Das M and Jog C J 1995 *Astrophys. J.* **451** 167

de Bernardis P *et al* 2000 *Nature* **404** 955

Dehnen W and Binney J 1998 *Mon. Not. R. Astron. Soc.* **294** 429

Downes D, Wilson T L, Bieging J and Wink J 1980 *Astron. Astrophys. Suppl.* **40** 379

Drimmel R 2000 *Astron. Astrophys.* **358** L13

Driver S P *et al* 1998 *Astrophys. J.* **496** L93

Englmaier P and Gerhard O 1999 *Mon. Not. R. Astron. Soc.* **304** 512

Evrard A E *et al* 2002 *Astrophys. J.* **573** 7

Ferguson H C, Dickinson M and Williams R 2000 *Annu. Rev. Astron. Astrophys.* **38** 667

Fernandez-Soto A *et al* 2001 *Astrophys. J. Suppl.* **135** 41

Ferrarese L and Merritt D 2000 *Astrophys. J.* **539** L9

Ferrerese L, Pogge R W and Peterson B M 2001 *Astrophys. J.* **555** L79

Figer D S *et al* 1999 *Astrophys. J.* **525** 750

Font A S, Navarro J F, Stadel J and Quinn T 2001 *Astrophys. J.* **563** L1

Freudenreich H T 1998 *Astrophys. J.* **492** 495

Frogel J A, Tiede G P and Kuchinski L E 1999 *Astrophys. J.* **117** 2296

Fux R 1997 *Astron. Astrophys.* **327** 983

——1999 *Astron. Astrophys.* **345** 787

Gebhardt K *et al* 2000a *Astrophys. J.* **543** L5

Gebhardt K, Bender R and Bower G 2000b *Astrophys. J.* **539** L13

Georgelin Y P and Georegelin Y M 1976 *Astron. Astrophys.* **49** 57

Ghigna S, Moore B and Governato F 2000 *Astrophys. J.* **544** 616

Giaconni R, Rosati P and Tozzi P 2001 *Astrophys. J.* **551** 624

Grabelsky D A, Cohen R S, Bronfman L and Thaddeus P 1988 *Astrophys. J.* **331** 181

Graham A W, Erwin P, Caon N and Trujillo I 2001 *Astrophys. J.* **563** L11

Gyuk G 1999 *Astrophys. J.* **510** 205

Haehnelt M G and Kauffmann G 2000 *Mon. Not. R. Astron. Soc.* **318** L35

Hanany S *et al* 2000 *Astrophys. J.* **545** L5

Hartmann D and Burton W B 1997 *Atlas of Galactic Neutral Hydrogen* (Cambridge: Cambridge University Press)

Heiligman G M 1987 *Astrophys. J.* **314** 747

Herrnstein J R, Moran J M and Greenhill L J 1999 *Nature* **400** 539

Ho L C 1999 *Observational Evidence for Black Holes in the Universe* ed S K Chakrabarti (Dordrecht: Kluwer) p 157

Holmberg J and Flynn C 2000 *Mon. Not. R. Astron. Soc.* **313** 209

Hoskin M 1985 *The Milky Way Galaxy (IAU Symp. 106)* ed H van Woerden *et al* p 11

Hughes D H *et al* 1998 *Nature* **394** 241

Hüttemeister S *et al* 1998 *Astron. Astrophys.* **334** 646

Hüttemeister S, Wilson T L, Bania T M and Martin-Pintado J 1993 *Astron. Astrophys.* **280** 255

Ibata R A, Gilmore G and Irwin M J 1994 *Nature* **370** 194

Jenkins A *et al* 1998 *Astrophys. J.* **499** 20

——2000 *Mon. Not. R. Astron. Soc.* **321** 372

Kaspi S *et al* 2000 *Astrophys. J.* **533** 631

Kauffmann G, Colberg J, Diaferio A and White S D M 1999 *Mon. Not. R. Astron. Soc.* **307** 529

Kauffmann G and Haehnelt M 2000 *Mon. Not. R. Astron. Soc.* **311** 576

Kenney J D P *et al* 1992 *Astrophys. J.* **395** L79

Kerr F J 1969 *Annu. Rev. Astron. Astrophys.* **7** 39

Klypin A, Kravtsov A, Valenzuela O and Prada F 1999 *Astrophys. J.* **522** 82

Kormendy J 1993 *The Nearest Active Galaxies* ed J Beckman, L Colina and H Netzer (Madrid: CSIC) p 197

Kormendy J and Richstone D O 1995 *Annu. Rev. Astron. Astrophys.* **33** 581

Kuijken K and Gilmore G 1991 *Astrophys. J.* **367** L9

La Rosa T N, Kassim N E, Lazio T, Joseph W and Hyman S D 2000 *Astrophys. J.* **119** 207

Lepine R D, Mishurov Y N and Dedikov S V 2001 *Astrophys. J.* **546** 234

Lin C C and Shu F H 1964 *Astrophys. J.* **140** 646

Lin C C, Yuan C and Shu F H 1969 *Astrophys. J.* **155** 721

Madau P, Ferguson H C and Dickinson M E 1996 *Mon. Not. R. Astron. Soc.* **283** 1388

Magorrian J *et al* 1998 *Astrophys. J.* **115** 2285

Malhotra S 1995 *Astrophys. J.* **448** 138

Martin-Pintado J, de Vicente P, Fuente A and Planesas P 1997 *Astrophys. J.* **482** L45

Merritt D and Ferrarese L 2001a *Astrophys. J.* **547** 140

——2001b *Mon. Not. R. Astron. Soc.* **320** 30

——2001c *The Central kpc of Starbursts and AGN (ASP Conference Series 249)* ed J H Knapen, J E Beckman, I Shlosman and T J Mahoney p 335

Merritt D, Ferrarese L and Joseph C L 2001 *Science* **293** 1116

Mihos J C and Hernquist L 1996 *Astrophys. J.* **464** 641

Milgrom M 1983 *Astrophys. J.* **270** 384

Miyoshi M *et al* 1995 *Nature* **373** 127

Möllenhoff C *et al* 1999 *Astron. Astrophys.* **352** L5

Moore B *et al* 1999 *Astrophys. J.* **524** L19

Moore B, Lake G and Katz N 1998 *Astrophys. J.* **495** 139

Morris M and Serabyn E 1996 *Annu. Rev. Astron. Astrophys.* **34** 645

Olling R P and Merrifield M R 2000 *Mon. Not. R. Astron. Soc.* **311** 361

——2001 *Mon. Not. R. Astron. Soc.* **326** 164

Paglione T A D, Jackson J M, Bolatto A D and Heyer M H 1998 *Astrophys. J.* **493** 680

Pascarelle S M *et al* 1996 *Astrophys. J.* **456** L21

Percival W J *et al* 2001 *Mon. Not. R. Astron. Soc.* **327** 1297

Piddington J H and Minnett H C 1951 *Aust. J. Sci. Res.* A **4** 459

Rodriguez-Fernandez N *et al* 2001 *Astron. Astrophys.* **365** 174

Rubin V C, Burstein D, Ford W K and Thonnard N 1985 *Astrophys. J.* **289** 81

Serjeant S B G, Eaton N and Oliver S J 1997 *Mon. Not. R. Astron. Soc.* **289** 457

Shapley H 1918 *Astrophys. J.* **48** 154

Shaver P A *et al* 1996 *Nature* **384** 439

——1999 *Highly Redshifted Radio Lines (ASP Conference Series 156)* ed C L Carilli, S J E Radford, K M Menten and G I Langston p 163

Shectman S A *et al* 1996 *Astrophys. J.* **470** 172

Silk J and Rees M J 1998 *Astron. Astrophys.* **331** L4
Sofue Y 1996 *Publications Astron. Soc. Japan* **49** 17
——1997 *Astrophys. J.* **458** 120
Solomon P M and Rivolo A R 1989 *Astrophys. J.* **339** 919
Springel V, Yoshida N and White S D M 2001 *New Astron.* **6** 51
Taylor J H and Cordes J M 1993 *Astrophys. J.* **411** 674
Tiede G P, Frogel J A and Terndrup P M 1995 *Astrophys. J.* **110** 2780
Toomre A 1981 *The Structure and Evolution of Normal Galaxies* ed S M Fall and
        D Lynden-Bell (Cambridge: Cambridge University Press)
Vallee J P 1995 *Astrophys. J.* **454** 119
van der Kruit P C 1984 *Astron. Astrophys.* **140** 470
van der Marel R P and Van den Bosch F C 1998 *Astrophys. J.* **116** 2220
Wandel A 1999 *Astrophys. J.* **519** L39
Wandel A, Peterson B M and Malkan M A 1999 *Astrophys. J.* **526** 579
Weaver H 1975a *Mercury* **8/9** 18
——1975b *Mercury* **11/12** 18
Weiner B J and Sellwood J A 1999 *Astrophys. J.* **524** 112
Whitford A E 1978 *Astrophys. J.* **226** 777
Xilouris E M *et al* 1999 *Astron. Astrophys.* **344** 868
Zinn R 1985 *Astrophys. J.* **293** 424

Chapter 3

The collapse to a black hole

Gernot Neugebauer
Friedrich-Schiller-Universität Jena, Germany

Now we have a broad view of the Milky Way, we turn back to the stars within the framework of general relativity. The purpose of this chapter is to review some of the astrophysical aspects of stellar black hole formation. When the nuclear fuel is exhausted, stars contract inwards under the influence of their own gravity. Our knowledge about the final stages of this collapse suggests that sufficiently massive stars inevitably leave black hole remnants, i.e. regions of spacetime in which gravity is so strong that neither matter nor light can ever escape. We discuss two collapse solutions of the Einstein equations. Whereas the Oppenheimer–Snyder model conveys the principal understanding of the dynamical transition to a (spherically symmetric) non-rotating black hole (collapse time, formation of the event horizon, communication problems of different observers), the parametric collapse of a rotating disk of dust suggests a preference for the extremely rotating (Kerr) black hole and a separation of spacetime domains.

3.1 Introduction

One of the most exciting predictions of Einstein's gravitational theory is that regions of spacetime in which gravity is so strong that neither matter nor light can ever escape exist. We owe this picture to a small number of stationary (axisymmetric) black hole solutions (Schwarzschild, Kerr, and Kerr–Newman solutions), whose sources consist of a physical singularity surrounded by an event horizon.

However, a large number of other stationary solutions with isolated singularities of completely different mathematical structure exist, so that we have to wonder whether and why only black holes should arise in reality. In this chapter, we will try to find an answer in the context of stellar evolution. As we will see, our understanding of the evolution processes now suggests that sufficiently

72

massive stars *inevitably* leave black hole remnants[1]. To see this, we need not discuss all steps of stellar evolution. It is sufficient to consider the final stage when nuclear fuel is exhausted and the star contracts slowly inwards under the influence of its own gravity thereby squeezing its pressure-sustaining electrons or photons onto its atomic nuclei. (From a macroscopic point of view its equation of state becomes softer.) The complex mathematical description of the resulting instability and the following implosion connected with mass ejection is beyond the scope of this chapter. Rather we will focus our attention on the question of whether the collapse can lead to stable stellar end products. If nuclear burning has finished, there is no thermal pressure balancing the star's self-gravitation. Hence the question is whether there are sufficiently high repulsion forces at all. Inspecting the different areas of physics we arrive at quantum physics and, with some knowledge of the structure of stars, at the repulsion forces of degenerate Fermi gases consisting of electrons or nucleons. The observation of white dwarfs and neutron stars confirms this picture.

White dwarfs can be thought to consist of two partial gases: free (highly degenerate) electrons and ionized nuclei forming a 'normal' gas. Only the degenerate electron gas will contribute a non-negligible amount to the pressure which supports the star against the gravitational forces. However, the massive nucleons must be considered to be the main source of gravitation.

Following Fowler [1] and Chandrasekhar [2], the quantitative (mathematical) description may start from a spherically symmetric static (non-rotating) fluid ball with the equation of state of the cold degenerate electron gas which can be written in a parametric form as

$$p = \frac{m_e c^2}{8\pi^2 \lambda_e^3} \left\{ x(1 + x^2)^{1/2} \left(\frac{2}{3} x^2 - 1 \right) + \ln[x + (1 + x^2)^{1/2}] \right\}$$

$$\rho \simeq \rho_0 = \frac{\mu_e m_B}{3\pi^2 \lambda_e^3} x^3 \qquad (x \text{ is a parameter})$$

connecting the isotropic (hydrostatic) pressure p and the mass density ρ, where ρ can be approximated by the mass density of the nucleons ρ_0. The constants m_e, m_B, c, and λ_e denote the electron mass, the baryonic mass, the velocity of light, and the de Broglie wavelength, respectively. The number of baryons per electron μ_e reflects the white dwarf's composition which depends on the fusion processes in the progenitor stars. (One of Chandrasekhar's original models uses $\mu_e = 2$ for white dwarfs consisting of helium; other compositions can be read off from figure 3.1). Once $p = p(\rho)$ is given, the spherically symmetric, static gravitational field equations can be integrated. A concise reformulation of the Einstein equation was given by Tolman, Oppenheimer, and Volkoff (TOV equations, see chapter 1, equations (1.89)–(1.92), this volume). It turns out though that Newtonian gravity is accurate enough to discuss white dwarf models.

---

[1] It should be mentioned, moreover, that black holes may also occur under other circumstances, e.g. as galactic or cosmological ('primordial') black holes.
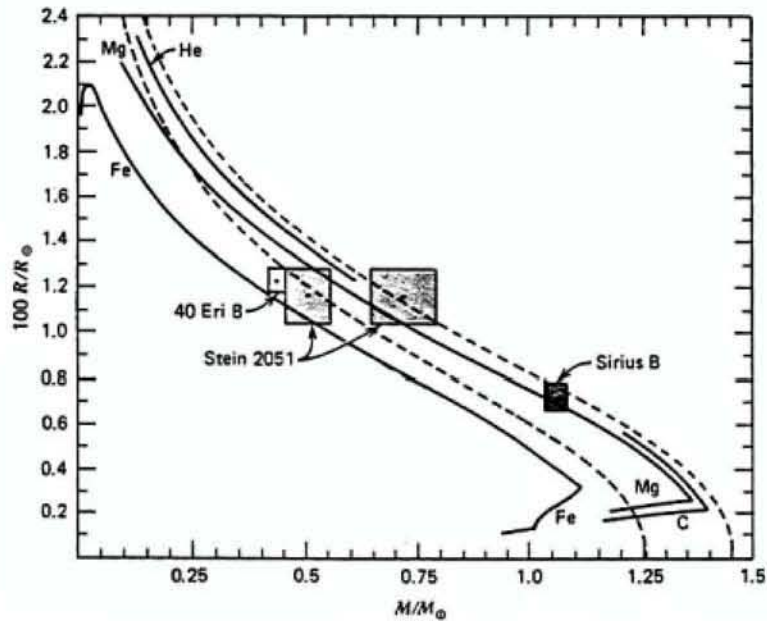
Figure 3.1. The relation between radius R and mass M for several white dwarf models. Depicted are curves for zero-temperature stars composed of helium, carbon, magnesium, and iron. The dashed curves denote Chandrasekhar's original models (the upper one for $\mu_e = 2$). There is a maximum mass of about 1.4 solar masses (adapted from Shapiro and Teukolsky [3]).

The mathematical problem consists in solving a set of ordinary differential equations for quantities depending on the radial coordinate r alone. One of these quantities is the mass density $\rho = \rho(r, R)$, where R denotes the radius of the star. Integrating over the volume, one obtains a relation between the mass M of the star and its radius R. Figure 3.1 adopted from the monograph of Shapiro and Teukolsky [3] shows several mass versus radius curves $R = R(M)$ for white dwarf models of different composition. The most remarkable result of the analysis as represented in figure 3.1 is the existence of a *maximum mass* of about 1.4 solar masses which cannot be exceeded by white dwarfs. This theoretical prediction has been confirmed by observation: all of the more than 1000 registered white dwarfs have masses smaller than 1.2 solar masses. In accordance with theoretical calculations their observed radii are planet-like ($R \sim 8000$ km).

Neutron star models were first worked out by Landau [4], Baade and Zwicky [5], and Oppenheimer and Volkoff [6]. They are based on the predominance of neutrons in the star matter after the characteristic collapse phases of the progenitor

star. The dominant effect is the mutual elimination of electrons and protons by inverse β-decay. Roughly speaking, a neutron star is a 'giant nucleus' formed by degenerate neutrons (10^{57} baryons/star) and held together by its huge self-gravity. The mathematical analysis of the spherically symmetric non-rotating (static) model follows the white dwarf analysis step by step. The main problem is to gain realistic equations of state (cf [7]) for the different zones of the star (central neutron lattice, neutron fluid, crust, cf [8]) from nuclear physics. Again, the mass density $\rho = \rho(r, R)$ and the mass–radius relation $M = M(R)$ can be calculated from the TOV equations (specified Einstein equations). Examples of mass versus radius curves calculated by means of different realistic equations of state are shown in figure 3.2 (adapted from [3]) and exhibit again that there are *maximum masses* which cannot be exceeded by neutron stars. Obviously, the characteristic values of these mass limits range from 1.4 to 3 solar masses; characteristic neutron star radii amount to 11–15 km. An inaccessible domain for any mass versus radius curve (gray domain in figure 3.2) independent of the equation of state ('Buchdahl limit') exists. This fact is a rigorous consequence of the Einstein equations for spherically symmetric static perfect fluid balls ('stars') which have no regular solution for[2] $R < 9/8 \times (2GM/c^2)$ (this can be inferred from the integration of the TOV equation, cf Stephani's textbook [9]). As we have seen in chapter 1 there is an explicit solution to the static spherically symmetric Einstein equations with constant mass density ρ ('Schwarzschild solution'). The corresponding mass versus radius relation $M = (4/3)\pi\rho R^3$ is sketched in figure 3.3 for 'nuclear densities'. The maximum masses as well as the corresponding radii of these simple models are a consequence of the principal fundamental relativistic limitation $R < 9/8 \times (2GM/c^2)$ and are not far from the realistic values of figure 3.2 ($\rho = 10^{15}$ g cm$^{-3} \rightarrow M \sim 3.6 M_\odot$, $R \sim 11.9$ km, $M_\odot$ is the solar mass). The model fails for lower mass densities for which a constant density throughout the star is not realistic.

The discovery of neutron stars is based on the pulsar mechanism: rapidly rotating neutron stars (periods of 10^{-3}–10^3 s) endowed with huge magnetic fields ($\sim 10^7$–10^9 Tesla) emit electromagnetic radiation (radio waves, X-rays) along the axis of the magnetic field. This radiation can be detected as a regular sequence of pulses which arrive at the terrestrial observer whenever the beam of the 'cosmic lighthouse' meets the earth. Astronomers know more than 800 radio pulsars and more than 700 X-ray pulsars and, in addition, about 10 binaries consisting of two neutron stars or combinations of neutron stars and white dwarfs. Most important for general relativity is the double star pulsar PSR 1913 + 16, a system consisting of two neutron stars of 1.4409 (± 5) and 1.3876 (± 5) solar masses, respectively (cf [10]). From the observed approach of the stars one can calculate the loss of mechanical energy. It corresponds exactly to the energy loss by gravitational radiation, as predicted by general relativity. This fact is considered to be the first (indirect) evidence for the existence of gravitational waves.

---

[2] G, gravitational constant; c, velocity of light.
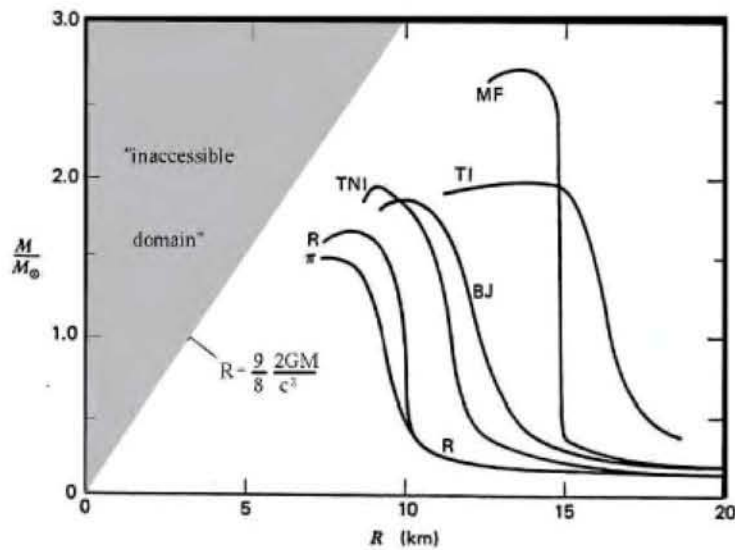
Figure 3.2. The relation between mass and radius for several neutron star models. None of the curves reflecting different equations of state exceeds a mass limit of about three solar masses (adapted from Shapiro and Teukolsky [3]). Einstein's theory itself forbids a penetration of mass–radius curves into the domain limited by the straight line $R = (9/4)(GM/c^2)$.

Let us summarize the quintessence of these introductory remarks: no physical mechanism is known to form end products of stellar evolution which exceed the maximum mass of about three solar masses. One might well ask whether there are mass ejection processes during the star evolution preventing the formation of more massive end products. For example, one could think of fundamental processes limiting the cores of supernova explosions to masses smaller than 3 solar masses. To the best of our knowledge such mechanisms do not exist. On the contrary, models of stellar evolution suggest that the 'total' collapse is inevitable if the initial stellar mass exceeds some critical value of about 80 solar masses.

As a consequence of Newton's gravitational theory, the inevitable collapse of a spherically symmetric mass distribution ends in a singular point (mass point). We will see in the next section that Einstein's theory foresees a more interesting fate; namely the formation of a black hole. It should be mentioned that a stellar collapse is merely one scenario for black hole formation. Accretion of gas by a white dwarf or a neutron star or the merging of two neutron stars (binaries) are considered to be further black hole forming processes. Moreover, primordial 'mini'-black holes ($M_{BH} \sim 10^{15}$ g) generated by perturbations in the early
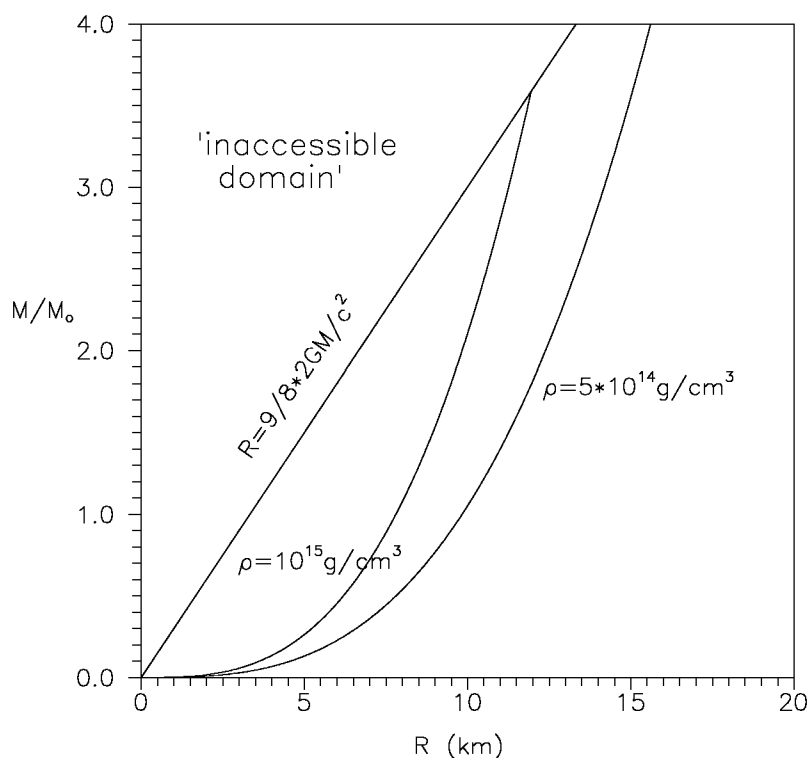
Figure 3.3. Mass versus radius relation for two 'Schwarzschild stars' with neutron star densities. Though unrealistic because of its constant mass density, the high density model provides a plausible maximum mass of about 3.6 solar masses.

Universe—and galactic black holes—come up for discussion.

3.2 Oppenheimer–Snyder collapse

3.2.1 Scenario and model

The Oppenheimer–Snyder model [11] describing the inevitable collapse of a ball of dust to a black hole is an exact solution of the general-relativistic field equations of gravitation. Rigorous solutions like this one are very important for deeper insight into the four-dimensional spacetime structure and enable us to ask correct physical questions in the context of general relativity. The point made here is to understand the formation of a black hole from 'normal' matter. Of course, dust is not very realistic in view of the complex structure of real collapsing stars.

Let us now model the main stages of a collapse:

(i)   Stationary phase before the collapse: the star could be thought to be a static spherically symmetric perfect fluid ball with the energy–momentum tensor

$$T_{ik} = (\rho + p)u_i u_k + p g_{ik} \tag{3.1}$$

where ρ, p, u_i, and g_{ik} are the mass density, the pressure, the four-velocity of the mass elements, and the metric respectively. (Units are chosen such that $G = c = 1$.) One could, e.g., think of the (interior and exterior) Schwarzschild solution, cf chapter 1).

(ii) End of nuclear burning and start of the collapse: all of a sudden, the pressure p breaks down ($p = 0$). For a moment, the resulting ball of dust ($p = 0$) with

$$T_{ik} = \rho u_i u_k \tag{3.2}$$

remains at rest.

(iii) Phase of collapse: since there is no pressure to balance gravity, the ball (star) begins to shrink. For dust, we expect an inevitable collapse and, finally, the formation of a black hole.

3.2.2   Solution of the field equations

3.2.2.1   The problem

To obtain the spacetime line element

$$ds^2 = g_{ik} \, dx^i \, dx^k \qquad (i, k = 0, 1, 2, 3;\ \text{summation convention}) \tag{3.3}$$

for the collapsing ball of dust, we have to integrate the Einstein equations

$$R_{ik} - \frac{R}{2} g_{ik} = 8\pi T_{ik} \tag{3.4}$$

where

$$\begin{aligned}
T_{ik} &= \rho u_i u_k \qquad \text{inside the ball (star)} \\
T_{ik} &= 0 \qquad \text{outside the ball (star).}
\end{aligned} \tag{3.5}$$

The solution must be asymptotically flat ($g_{ik} \rightarrow \eta_{ik}$ at infinity, the 'boundary value problem') and regular everywhere for all times $t \geq 0$. According to our model, the star should be at rest for $t = 0$ (the 'initial value problem'). To solve the boundary/initial value problem, we will solve the interior and exterior field equations separately and match the interior to the exterior part afterwards ('matching problem'). The procedure relies heavily on spherical symmetry.

3.2.2.2   Exterior solution

An aid to the understanding of the following material can be gained by reading the appropriate passages of the excellent book from Misner, Thorne and Wheeler [12] (known as MTW).

Spherical symmetry means that ds^2 in (1.3) is form invariant under the group of spatial orthogonal transformations and can therefore be put in the form

$$ds^2 = g_{AB}dx^A dx^B + r^2 d\Omega^2$$
$$d\Omega^2 := d\vartheta^2 + \sin^2\vartheta \, d\varphi^2 \qquad (A, B = 0, 1) \tag{3.6}$$

where (g_{AB}) and r are functions of x^0 and x^1 alone,

$$g_{AB} = g_{AB}(x^0, x^1) \qquad r = r(x^0, x^1) \tag{3.7}$$

and $x^2 = \vartheta \, (0 \leq \vartheta \leq \pi)$ and $x^3 = \varphi \, (0 \leq \varphi \leq 2\pi)$ are the usual spherical coordinates.

The line element (3.6) may further be reduced by a suitable choice of the coordinates (x^0, x^1). The following three examples are useful for collapse discussions:

(i) *Schwarzschild coordinates*:

$$ds^2 = -e^{\nu(t,r)} \, dt^2 + e^{\lambda(t,r)} \, dr^2 + r^2 \, d\Omega^2. \tag{3.8}$$

This form may be obtained by a coordinate transformation $r = r(x^0, x^1)$, $t = (x^0, x^1)$.

The vacuum equations (3.4), (3.5),

$$R_{ik} = 0 \tag{3.9}$$

give $\partial\lambda/\partial t = 0$, $\partial^2\nu/\partial t^2 = 0$, first of all, and after a time scaling of the form $t' = t'(t)$ the exterior Schwarzschild solution

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \frac{dr^2}{1 - 2M/r} + r^2 \, d\Omega^2 \tag{3.10}$$

where M is the reduced mass of the star. (Note that we have put $G = 1, c = 1$. Hence we do not distinguish between the mass M measured in kilograms and the reduced mass MG/c^2 expressed in metres.) Obviously, the only vacuum solution with spherical symmetry is the exterior Schwarzschild solution[3] (Birkhoff's theorem, 1923) and that solution is static. As a consequence, a collapsing spherically symmetric star cannot emit gravitational waves. The Schwarzschild metric (3.10) is regular for $r > 2M$ and has, apparently, a singularity at the

---

[3] MTW formulation, cf MTW [12]: let the geometry of a given region of spacetime (i) be spherically symmetric and (ii) be a solution to the Einstein equations in vacuum. Then that geometry is necessarily a piece of the Schwarzschild geometry.

Schwarzschild radius $r = 2M$. However, the geometry (the curvature tensor) is regular there. Thus, 'better' coordinates in the domain $r \leq 2M$ must exist.

(ii) *Kruskal–Szekeres coordinates*:

$$ds^2 = b(du^2 - dv^2) + r^2\, d\Omega^2, b = b(u, v), r = r(u, v). \tag{3.11}$$

Any two-metric (such as g_{AB} in (3.6)) is conformally equivalent to a flat two-metric. We have put $x^0 = u$, $x^1 = v$. Inserting (3.11) into the field equations (3.9), we obtain $b = b(u, v)$ and $r = r(u, v)$,

$$ds^2 = \frac{32M^3}{r} e^{-r/2M} \left(du^2 - dv^2\right) + r^2 d\Omega^2$$

$$\left(\frac{r}{2M} - 1\right) e^{r/2M} = u^2 - v^2. \tag{3.12}$$

The Kruskal–Szekeres coordinates (u, v) are related to the Schwarzschild coordinates by

$$u = \left(\frac{r}{2M} - 1\right)^{1/2} e^{r/4M} \cosh\left(\frac{t}{4M}\right)$$

$$v = \left(\frac{r}{2M} - 1\right)^{1/2} e^{r/4M} \sinh\left(\frac{t}{4M}\right) \quad \text{when } r > 2M$$

$$u = \left(1 - \frac{r}{2M}\right)^{1/2} e^{r/4M} \sinh\left(\frac{t}{4M}\right)$$

$$v = \left(1 - \frac{r}{2M}\right)^{1/2} e^{r/4M} \cosh\left(\frac{t}{4M}\right) \quad \text{when } r < 2M. \tag{3.13}$$

There is no coordinate singularity at $r = 2M$. Kruskal–Szekeres coordinates are well suited for the global analysis of black holes.

(iii) *Ingoing Eddington–Finkelstein coordinates*:

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dV^2 + 2dV\, dr + r^2\, d\Omega^2. \tag{3.14}$$

This line element can be obtained from (3.10) by the coordinate transformation

$$r = r \qquad V = t + r + 2M \ln\left|\frac{r}{2M} - 1\right|. \tag{3.15}$$

The coordinates are adapted to ingoing light rays $V = $ constant, cf MTW [12].

3.2.2.3   Interior solution

The field equations (3.4) together with the dust matter model (3.5) imply local mass conservation

$$(\rho u^i)_{;i} = 0 \tag{3.16}$$

and geodesic motion of the mass elements,

$$\frac{Du^i}{D\tau} = u^i{}_{;k}u^k = 0. \tag{3.17}$$

Equations (3.16) and (3.17) are a direct consequence of the local energy–momentum conservation,

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

To get a manageable model, we are on the look-out for further hydrodynamic simplifications. Because of the radial motion of the mass elements, the *vorticity* of the four-velocity has to vanish. Let us assume, in addition, that the motion be shear free. Then we arrive at [13],

$$u_{i,k} = \tfrac{1}{3}u^m{}_{;m}(g_{ik} + u_i u_k). \tag{3.18}$$

(Note that (3.18) implies $u_{i,k} - u_{k,i} = 0$ whence $Du^i/D\tau = 0$.) We may now choose the world lines of the mass elements to be the time lines ('*comoving coordinates*'),

$$(u^i) = (1, 0, 0, 0). \tag{3.19}$$

Then, by means of (3.18), the line element (3.3) can be put in the following form

$$ds^2 = -dt^2 + a^2(t, x^1, x^2, x^3)g_{ab}\,dx^a\,dx^b \tag{3.20}$$

where g_{ab} does not depend on t.

As a consequence of the field equations (3.4) for dust (3.5), the subspace $t = $ constant has a constant (vanishing, positive or negative) curvature. Introducing spherical coordinates χ $(0 \le \chi < \chi_0)$, ϑ $(0 \le \vartheta \le \pi)$, φ $(0 \le \varphi \le 2\pi)$, the line element (3.20) takes the form

$$ds^2 = -dt^2 + a^2(t)(d\chi^2 + f(\chi)[d\vartheta^2 + \sin^2\vartheta\,d\varphi^2])$$
$$f = (\chi^2, \sin^2\chi, \sinh^2\chi) \tag{3.21}$$

where f characterizes the three curvature types. The remaining field equations (and their consequence (3.16)) reduce to the simple first-order system

$$3(\dot{a}^2 + \varepsilon) = 8\pi a^2\rho \qquad \frac{d}{dt}(\rho a^3) = 0 \tag{3.22}$$

where $\varepsilon = (0, 1, -1)$ corresponds to the three values of f, and a dot denotes time derivative.

The line element (3.21) describes the interior geometry of an expanding ($\dot{a} > 0$) or contracting ($\dot{a} < 0$) dust ball. Its three-surface $\chi = \chi_0$ is the interface to the exterior Schwarzschild region. Interestingly, our initial value problem ('star at rest'),

$$t = 0: a(0) = a_m \qquad \dot{a}(0) = 0 \tag{3.23}$$

singles out positive curvature,

$$\varepsilon = 1. \tag{3.24}$$

($\varepsilon = 0, -1$ is not compatible with $\dot{a}(0) = 0$ in (3.22).)

Then the solution of (3.22)–(3.24) is given by

$$a = \frac{a_m}{2}(1 + \cos \eta) \qquad t = \frac{a_m}{2}(\eta + \sin \eta) \qquad \rho = \frac{3a_m}{8\pi}a^{-3} \tag{3.25}$$

where

$$0 \le \eta \le \pi.$$

The equations (3.25) describe a dust ball whose scale factor $a(t)$ shrinks from the initial value $a = a_m(\eta = 0, t = 0)$ to zero ($\eta = \pi, t = \pi a_m/2$).

The dust model has a fundamental meaning for cosmology: the equations (3.21), (3.22) describe the geometry and mass distribution of an open ($\varepsilon = 0, -1$) or closed ($\varepsilon = 1$) Friedman universe.

We have now to match the interior (Friedman) (3.21), (3.25) and the exterior (Schwarzschild) solution.

3.2.2.4   Matching of geometries

The interface between the interior and exterior geometry is formed by the geodesic (world) lines of the mass elements on the surface of the dust ball. This 'world tube' can be described in the interior as well as in the exterior coordinates. In interior coordinates, the radial surface geodesics are characterized by the equation $x^0 = \tau$ (cf (3.19), $dx^0/d\tau = 1$) and constant values of ϑ, φ, χ for all proper times $\tau \ge 0$,

$$\tau \ge 0 : x^0 = \tau, \vartheta = \vartheta_0 \qquad (0 \le \vartheta_0 \le \pi)$$

$$\chi = \chi_0, \varphi = \varphi_0 \qquad (0 \le \varphi_0 \le 2\pi). \tag{3.26}$$

However, the same world lines have to be radial geodesics of the exterior Schwarzschild solution (3.10). Integrating the underlying equation

$$\frac{d^2x^i}{d\tau^2} + \Gamma^i_{kl}\frac{dx^k}{d\tau}\frac{dx^l}{d\tau} = 0 \tag{3.27}$$

for the initial values

$$\tau = 0 : r(\tau = 0) = r_0, \dot{r}(\tau = 0) = 0 \ (\vartheta = \vartheta_0, \varphi = \varphi_0) \tag{3.28}$$

we obtain

$$\tau \ge 0 \ (\eta \ge 0) :$$

$$r = \frac{r_0}{2}(1 + \cos \lambda) \qquad \tau = \sqrt{\frac{r_0^3}{8M}}(\lambda + \sin \lambda)$$

$$t = 2M \ln \left| \frac{\sqrt{r_0/2M - 1} + \tan \lambda/2}{\sqrt{r_0/2M - 1} - \tan \lambda/2} \right| + 2M\sqrt{\frac{r_0}{2M} - 1}\left[\lambda + \frac{r_0}{4M}(\lambda + \sin \lambda)\right] \tag{3.29}$$

where the real parameter λ is restricted by

$$0 \le \lambda \le \pi. \tag{3.30}$$

Obviously, $\tau = 0$ corresponds to the parameter value $\eta = 0$.

The identification of the equations (3.26) and (3.29) requires *coordinate-free* criteria. It can easily be seen that the proper time τ,

$$d\tau^2 = -ds^2 = -g_{AB}\, dx^A\, dx^B - r^2(x^0, x^1)(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2)$$
$$x^0 = t \tag{3.31}$$

and the circumference $C(\tau)$ of the three-dimensional interface ('world tube') along the slice $S : x^A = \text{constant}\ (A = 0, 1)$, $\vartheta = \frac{\pi}{2}$,

$$C(\tau) = \int_S ds = \int_0^{2\pi} r(x^0, x^1)\, d\varphi = 2\pi r(x^0, x^1) \tag{3.32}$$

have an invariant geometrical meaning. Making use of the equations (3.21), (3.25), and (3.26), we obtain in interior coordinates

$$\tau = t = \frac{a_m}{2}(\eta + \sin\eta) \qquad (0 \le \eta \le \pi)$$
$$C(\tau) = \pi a_m(1 + \cos\eta)\sin\chi_0 \tag{3.33}$$

whereas, according to (3.29), the calculation in exterior coordinates yields

$$\tau = \sqrt{\frac{r_0^3}{8M}}(\lambda + \sin\lambda)$$
$$C = \pi r_0(1 + \cos\lambda) \qquad (0 \le \lambda \le \pi). \tag{3.34}$$

From the identification of the equations (3.33) and (3.34) we have

$$\lambda = \eta \qquad r_0 = a_m \sin\chi_0 \qquad M = \frac{a_m}{2}\sin^3\chi_0 \qquad (0 \le \eta \le \pi). \tag{3.35}$$

Thus the 'interior' constants a_m and χ_0, describing the maximum extension of the star $(a = a_m)$ and its surface $(\chi = \chi_0)$, can be expressed in terms of the star's total mass M and its maximum 'radius' r_0.

It can be shown that the (necessary) conditions (3.35) are sufficient to glue together smoothly the pieces of Friedman geometry describing the star's interior,

$$ds^2 = -dt^2 + a^2(t)[d\chi^2 + \sin^2\chi\, d\Omega^2] \tag{3.36}$$
$$a(t) \text{ as in (3.25)} \qquad 0 \le \chi \le \chi_0$$

and Schwarzschild geometry describing its exterior,

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \frac{dr^2}{1 - 2M/r} + r^2 d\Omega^2 \tag{3.37}$$
$$r(t) \le r < \infty \qquad r(t) \text{ as in (3.29)}.$$

Both pieces of geometry glued together smoothly along the interface form only one solution to the Einstein equations. It describes a collapsing model star whose circumference shrinks permanently from $C = 2\pi r_0$ ($\lambda = 0$, 'star at rest—start of collapse') to $C = 0$ ($\lambda = \pi$, 'end of collapse').

3.2.3 Physical interpretation

According to (3.36), there is no obvious geometrical singularity during the collapse, not even when a mass point of the star's surface reaches the 'dangerous' coordinate value $r = 2M$ in (3.37). This is a clear hint to change the coordinate system (3.37) and to introduce 'better' exterior coordinates, as, e.g., Kruskal–Szekeres or Eddington–Finkelstein coordinates.

To obtain the world line of a surface point of the star in Eddington–Finkelstein coordinates, one has to insert the surface geodesics (3.29) into the coordinate transformation (3.15),

$$r(\lambda) = \frac{r_0}{2}(1 + \cos \lambda) \qquad \tau = \sqrt{\frac{r_0^3}{8M}}(\lambda + \sin \lambda)$$

$$V(\lambda) = r(\lambda) + 2M \ln \left| \frac{r(\lambda)}{2M} - 1 \right| + 2M \ln \left| \frac{\sqrt{r_0/2M - 1} + \tan \lambda/2}{\sqrt{r_0/2M - 1} - \tan \lambda/2} \right|$$

$$+ 2M \sqrt{\frac{r_0}{2M} - 1} \left[\lambda + \frac{r_0}{4M}(\lambda + \sin \lambda) \right] \qquad (0 \leq \lambda \leq \pi).$$

$$(3.38)$$

The diagram shown in figure 4 was taken from MTW [12, p 849]. It uses slightly modified Eddington–Finkelstein coordinates

$$\tilde{V} = V - r \qquad \tilde{r} = r \tag{3.39}$$

and illustrates the collapse of a star whose radius r drops from $r_0 = 10M$ to $r = 0$. According to (3.38), the proper time interval $\Delta \tau$ of the collapse, as measured by a comoving observer placed on the star's surface, is surprisingly short[4]

$$\Delta \tau = \sqrt{\frac{(10M)^3}{8M}}\pi \sim 35.1M. \tag{3.40}$$

Even the critical radius $r = 2M$ is reached in a similarly short proper time of $\Delta \tau = 33.7M$. Since, as we will see, this event marks the birth of a black hole, an external observer placed at fixed spatial coordinate values (at $r = 10M$ in figure 3.4) could ask him- or herself after a while whether the black hole had

---

[4] The corresponding proper time expressed in seconds for a dust ball with the radius and the mass of the sun is $\Delta \tau = \frac{1}{c} \sqrt{\frac{(0,696)^3 \cdot 10^{18}}{11,84}} \pi$ km $\simeq 1768$ s ($c = 2.997 \times 10^5$ km s^{-1}).
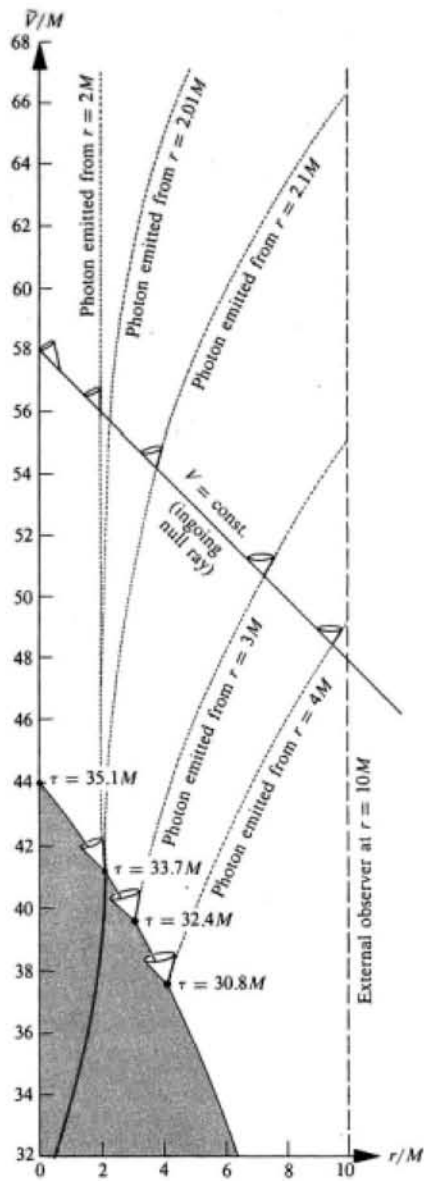
Figure 3.4. Oppenheimer–Snyder collapse in modified Eddington–Finkelstein coordinates (adapted from MTW [12]). The diagram depicts a series of photons emitted radially from the surface of the collapsing star and received by an observer at $r = r_0 = 10M$. Any photon emitted radially at the Schwarzschild radius $r = 2M$ stays at $r = 2M$ forever. This external *event horizon* is the continuation of the internal *event horizon* (full curve in the shaded interior region of the star).

already formed or not. The scenario sketched in figure 3.4 demonstrates that such a question is ill posed. To communicate with each other, the comoving observer and the external observer could exchange light signals (photons). Since the Eddington–Finkelstein coordinates (3.14) are adapted to radial ingoing light rays (V, ϑ, φ : constants, $ds^2 = 0$), the world line of an ingoing photon in the modified coordinates (3.39) is given by

$$ds^2 = 0 : \qquad \tilde{V} = V - r \qquad V = \text{constant.} \tag{3.41}$$

All ingoing null rays are parallel to the one sketched in figure 3.4, i.e. all photons emitted from the external observer reach the domain $r = 0$. According to (3.14), we get for the outgoing light rays describing photons emitted from the surface of the collapsing star

$$ds^2 = 0 : \qquad \tilde{V}(r) = V - r = r + 4M \ln \left| \frac{r}{2M} - 1 \right| + D \qquad D = \text{constant.} \tag{3.42}$$

One may fix the constant $D = D(\eta)$ from the intersection with the star's surface (3.38). Figure 3.4 depicts a series of such photon world lines emitted from the surface of the collapsing star and received by the observer at $r = 10M$. It illustrates the way in which the external observer distinguishes the collapse. Consider a series of uniformly spaced light signals emitted by the comoving observer. Received at the external observer, they would become more and more widely spaced. Finally, any photon emitted at $r = 2M$ stays at $r = 2M$ forever and can never reach the external observer. For this reason, the domain $r = 2M$ is called the *event horizon*. In the diagram it is marked by the world line of the photon emitted from $r = 2M$. After the Schwarzschild radius $r = 2M$ has been passed, the external observer never receives a signal: all photons emitted from the star's surface are sucked into the singularity at $r = 0$, which is a real singularity of spacetime geometry.

Let us now return to the question of black hole formation. From the point of view of the external observer, the star never gets beyond its Schwarzschild radius $r = 2M$. This seems to contradict the observation of the comoving observer that the star rapidly reaches $r = 2M$ and $r = 0$. A physicist, who applies Einstein's theory, must not identify his or her position with that of the external or the comoving observer. The physicist knows the geometry of the four-dimensional spacetime, the world lines of matter and observers, and the physical processes in spacetime. He can visualize this knowledge by four-dimensional diagrams such as that in figure 3.4. This complete information enables him/her to formulate and answer 'reasonable' (i.e. physical) questions. From his or her four-dimensional point of view, the event 'star surface passes the Schwarzschild radius $r = 2M$' marks the formation of a black hole characterized by its event horizon. He or she knows that processes of the black hole's history, such as the emission of photons from the star's surface, can influence the remotest future of an external observer.

The four-dimensional view of the gravitational phenomena connected with the black hole formation requires the analysis of the *internal* geometry of the black hole and the processes *inside* the black holes, too. The internal part of spacetime cannot be ignored with the argument that the external observers (as we are) will never obtain information from the interior.

The Eddington–Finkelstein coordinates are adapted to the vacuum domain outside the star. To depict the world lines of the mass elements or to extend the event horizon to the interior of the star (gray domain in figure 3.4), one needs 'internal' Eddington–Finkelstein coordinates. The extension is not unique. An obvious choice is to use in (3.36) radially ingoing light rays[5] $V = $ constant and the radial coordinate $r = a(t) \sin \chi$.

$$V(\chi, t) = F(\chi + \eta(t)) \qquad t = \frac{a_m}{2}(\eta + \sin \eta)$$

$$r = r(\chi, t) = a(t)\sin \chi \qquad a(t) = \frac{a_m}{2}(1 + \cos \eta(t)) \qquad (3.43)$$

where F is an arbitrary function of $\chi + \eta$. F can be fixed by the identification of the internal ingoing light rays $V = F(\chi + \eta) = $ constant with the external light rays $V = $ constant along the surface.

The horizon inside the star consists of those outgoing geodesic lines that meet the external event horizon at the star's surface $\chi = \chi_0$,

$$V = F(2\chi - 3\chi_0 + \pi)$$

$$r = \frac{a_m}{2}(1 + \cos[\chi - 3\chi_0 + \pi])\sin \chi \qquad (3.44)$$

$$(0 \leq \chi \leq \chi_0).$$

In figure 3.4, the internal event horizon is depicted by the full line inside the star.

All internal ingoing light rays 'beneath' that line arrive at the surface before the black hole has passed through its Schwarzschild radius $r = 2M$ and can therefore escape to infinity. The light rays beyond the line arrive at the surface inside the black hole and get pulled into the singularity at $r = 0$.

Unfortunately, there is not enough space for a detailed discussion of further effects in this chapter. Following the textbook of MTW [12], I confine myself to a description of some interesting results.

Light propagation. Light emitted from the collapsing star becomes more and more redshifted and reaches a relative redshift limit of $z \simeq 2$ at a distant external observer. The total luminosity decays exponentially in time. Light from the star in its late stages (before the black hole transition) is not dominated by 'radial photons' but by photons that were deposited by the star in (unstable) circular orbits as its surface passed through $r = 3M$.

Fate of the star beyond its horizon. The star and the entire internal geometry collapse inevitably in a spacetime singularity which crushes the collapsing matter to infinite density.

---

[5] Insert V in (3.36) in order to verify that $V = $ constant describes ingoing light rays.

Fate of the comoving observer. There are no special problems at $r = 2M$ (imagine the 'soft' formation of a huge galactic black hole of 10^6 solar masses!) Eventually, the observer crushes to zero volume and indefinitely extended length at $r = 0$.

The Tolman–Oppenheimer model has been generalized for other dust balls [14, 15] and different energy–momentum tensors [16]. Special attention is being paid to the non-spherical collapse as a source of gravitational radiation. There is justified hope that gravitational waves emitted from type IIa supernovae and other collapsing cosmic sources could be detected with the aid of the laser interferometers of the VIRGO, LIGO, or GEO 600 projects.

Most insiders are convinced that the end product of a non-spherical collapse with non-vanishing angular momentum is a rotating black hole whose external geometry is described by the stationary and axially symmetric Kerr solution (see Hawking and Ellis [17] for a review) which, in turn, is determined uniquely by the parameters of mass and angular momentum. It would be highly desirable to elaborate more precisely on such an assertion. Before that, the question remains undecided as to whether cosmic collapse processes lead inevitably to the formation of black holes. Another very important problem is the question whether the singularity at the end point of spherical collapse is typical for all collapse processes or whether it can be removed in more general collapses. Such questions were first analyzed by Penrose [18] and Hawking and Penrose [19]. For an introduction to the analysis of singularities one should study the book of Hawking and Ellis [17].

3.3   Rotating matter and black hole formation

Unfortunately, the scenario described in the preceding section cannot be extended without difficulty to matter with angular momentum. Though we do not possess a dynamical model for collapsing matter, we expect some insight into the process of the formation of rotating black holes by the discussion of an 'adiabatic' transition. For this purpose, we consider a stationary and axisymmetric global solution to Einstein's equations describing a rigidly (uniformly) rotating disk of dust. The solution allows a 'parametric' collapse, i.e. it has a black hole limit for characteristic values of its parameters. In the following we present some details of the solution which has been found by Neugebauer and Meinel [20] as the solution of a boundary value problem first formulated and approximately solved by Bardeen and Wagoner [21, 22] (see [23] for a review).

The line element of any stationary and axially symmetric solution for isolated and uniformly rotating sources can be cast into the form

$$ds^2 = -e^{2U}(dt + a\, d\varphi)^2 + e^{-2U}(e^{2k}[dr^2 + dz^2] + W^2\, d\varphi^2) \qquad (3.45)$$

where r, z, φ are cylindrical coordinates. The Newtonian gravitational potential U, the gravitomagnetic potential a, the superpotential k, and the azimuthal
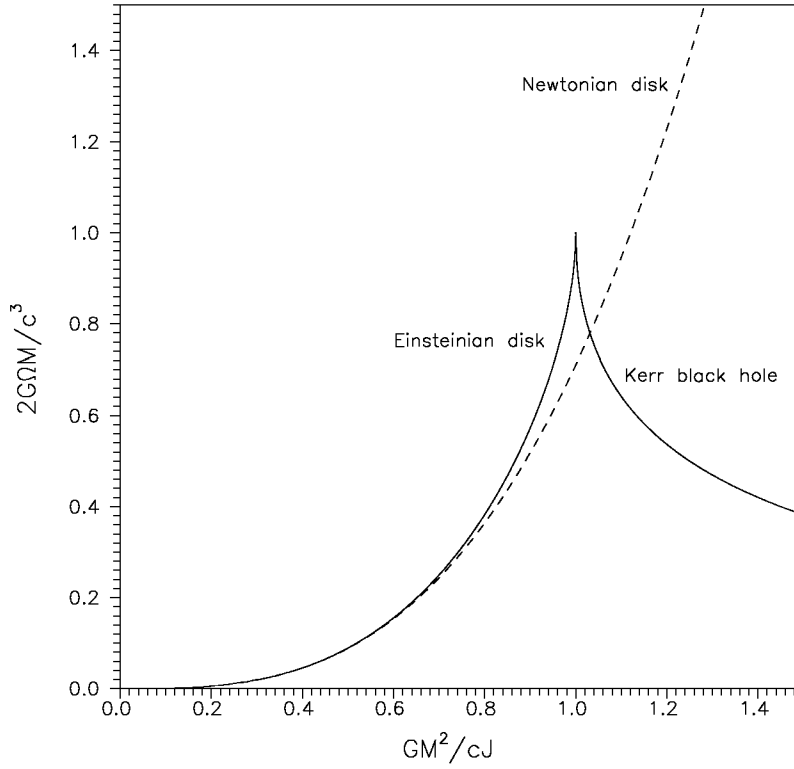
Figure 3.5. Relation between ΩM and M^2/J for the Newtonian Maclaurin disk (broken curve), the general-relativistic dust disk and the Kerr black hole. The plot illustrates the phase transition from the disk to the black hole at $M^2 = J$.

potential W depend on r and z alone, i.e. the metric (3.45) allows an Abelian group of motions G_2 with the generators (Killing vectors) ξ^i, η^i,

$$\xi^i = \delta^i_t, \quad \xi^i \xi_i < 0 \qquad \text{stationarity}$$
$$\eta^i = \delta^i_\varphi, \quad \eta^i \eta_i > 0 \qquad \text{axisymmetry.} \tag{3.46}$$

Outside matter, one may choose $W = r$.

Rotational motion of the matter about the symmetry axis z means

$$u_i = e^{-V}(\xi_i + \Omega \eta_i) \qquad u_i u^i = -1 \tag{3.47}$$

i.e. the four-velocity u_i is a linear combination of the Killing vectors. For rigidly rotating bodies, such as our disk of dust, the angular velocity is a constant

$$\Omega = \text{constant.} \tag{3.48}$$
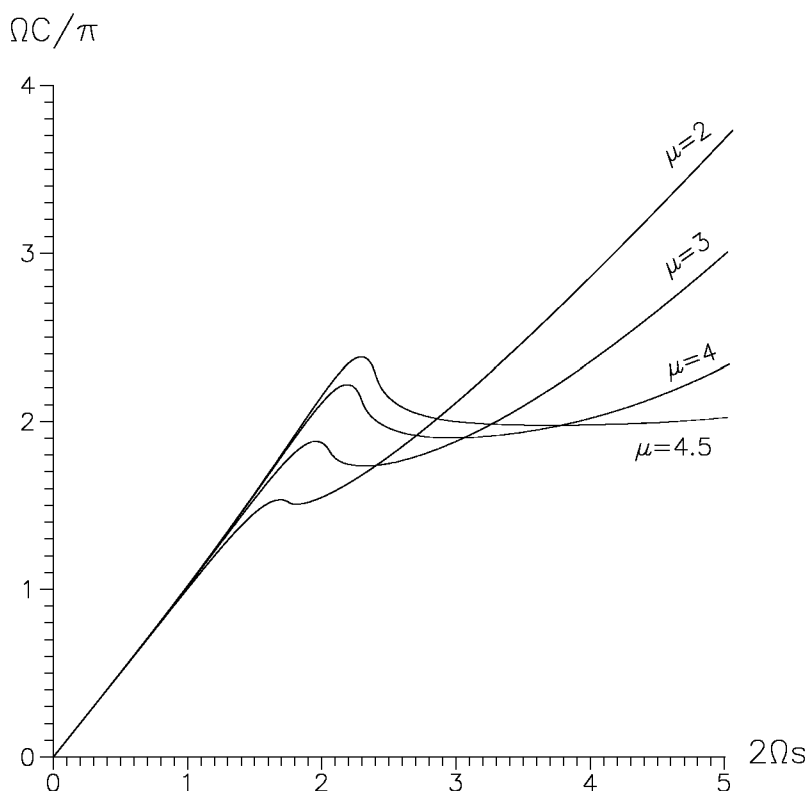
Figure 3.6. Geometry in the disk plane. The circumferential diameter C/π of a circle around the centre of the disk versus the real distances from the centre for increasing values of μ (here $\Omega C/\pi$ and Ωs are dimensionless quantities, $c = 1$) are depicted.

The task of solving the Einstein equations with the energy–momentum tensor (3.5) for an infinitesimally thin rigidly rotating disk of dust leads to the following boundary value problem: find a regular solution of Einstein's vacuum equations in the form (3.45), where

(i) the line element (3.45) becomes Minkowskian at infinity,

$$r^2 + z^2 \rightarrow \infty: \qquad ds^2 = -dt^2 + dr^2 + dz^2 + r^2 d\varphi^2 \qquad (3.49)$$

and

(ii) the co-rotating potentials $U' = V$ and a', as measured by an observer comoving with the disk, satisfy the conditions

$$z = 0, 0 \le r \le r_0: \qquad U' = V_0 = \text{constant}, \frac{\partial a'}{\partial z} = 0 \qquad (3.50)$$

$\Omega C/\pi$
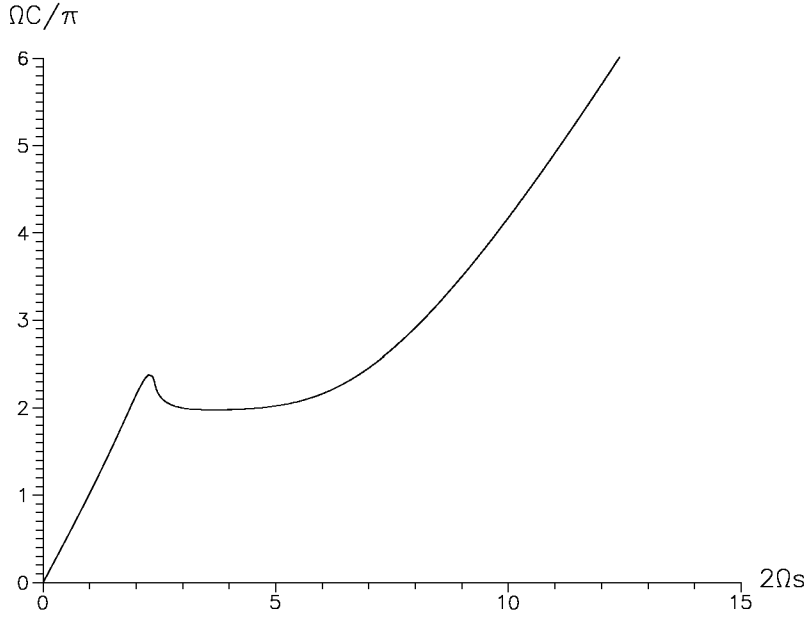
Figure 3.7. For ultrarelativistic values of μ (here $\mu = 4.5$), the 'interior region' around the disk (around the local maximum on the left-hand side) is far from the 'exterior region' (right-hand ascending branch of the curve), which becomes more and more Kerr-like.

in the disk (r_0 is the coordinate radius of the disk).

Equations (3.47) and (3.46) imply

$$e^{2U'} = e^{2U}[(1 + \Omega a)^2 - r^2\Omega^2 e^{-4U}] \tag{3.51}$$

which means that the boundary conditions contain two constant parameters, V_0 and Ω. Consequently, the disk of dust solution is a two-parameter solution. Unfortunately, there is not enough space (or time) to write out the solution explicitly and to prove the facts on which the following description is based. Readers are referred to the original papers [20–23].

An impression of the 'parametric' collapse of the disk can be gained from the parameter relations

$$V_0 = V_0(M, J) \qquad \Omega = \Omega(M, J) \tag{3.52}$$

which connect the disk parameters V_0 and Ω with the far-field parameters (energy-mass) M and angular momentum J [23]. Figure 3.5 combines the parameter relations between ΩM and M^2/J for the classical Maclaurin disk (broken curve), the general-relativistic disk, and the Kerr black hole. It indicates a phase transition from the disk to the extreme Kerr black hole at $M^2/J = 1$ and $2\Omega M = 1$. This
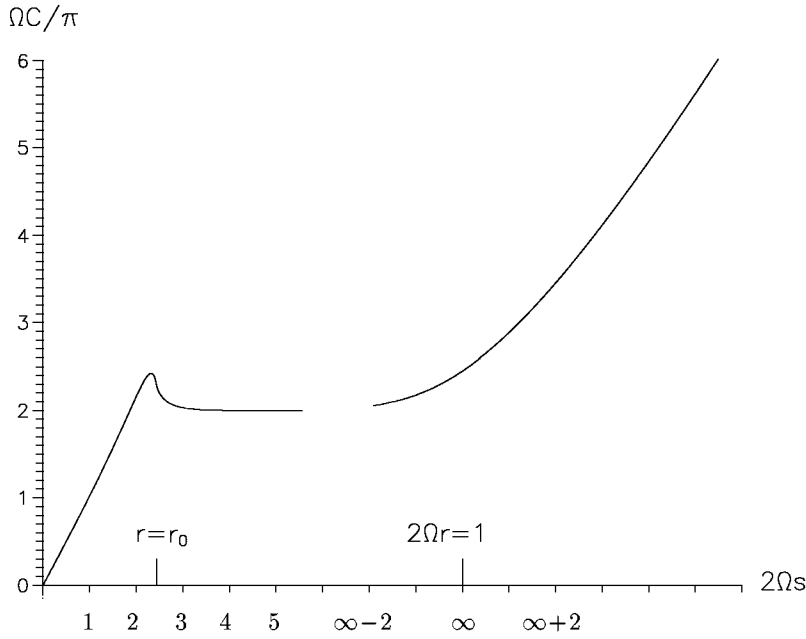
Figure 3.8. In the limit $\mu = \mu_0$, the 'disk world' (left-hand branch) and the 'world of the extreme Kerr black hole' (right-hand branch) are separated from each other. The point labelled ∞ on the abscissa corresponds to a coordinate radius $r = 1/2\Omega$. Points of the 'Kerr world' (right-hand branch) are at infinite distance from the disk (in the left-hand branch).

behaviour of rotating matter is completely different from that of static spherically symmetric fluid spheres which have to obey the restriction $R > (9/4)(GM/c^2)$ (R, radius; M, mass of the sphere) and can not therefore occupy the infinitesimal neighbourhood $(2GM/c^2 \leq R \leq (9/4)(GM/c^2))$ of their black hole state at $R = 2(GM/c^2)$ (see [9]). Figures 3.6–3.8, which correspond to the original figures in [22], convey a more detailed picture of the parametric black hole transition ('collapse') of the disk. They show the circumference C of a circle in the disk plane (cf (3.32)) about the centre of the disk versus the real distance s from the centre,

$$s = \int_0^r dr' \exp[k(0, r) - U(0, r)] \tag{3.53}$$

for different values of the parameter μ,

$$\mu = 2\Omega^2 r_0^2 e^{-2V_0} \qquad 0 \leq \mu \leq \mu_0 = 4.62966\ldots.$$

The limits $\mu \to 0$ and $\mu \to \mu_0$ mark the non-relativistic Maclaurin disk and the ultra-relativistic limit, respectively. For increasing values of μ, figures 3.6–3.8

illustrate a separation process for two domains of the disk solution which results, as indicated in figure 3.8, in a disk world (left-hand branch, limit 1) at an infinite distance from the world of an extreme ($M^2 = J$) Kerr black hole (right-hand branch, limit 2). A detailed analysis of the complete solution confirms the result: for $\mu \rightarrow \mu_0$, there are two limits of the disk spacetime. Whether or not this phenomenon can also be found for a *dynamical* collapse must remain the fruit of future work.

References

[1] Fowler R H 1926 Dense matter *Mon. Not. R. Astron. Soc.* **87** 114
[2] Chandrasekhar S 1931a The density of white dwarf stars *Phil. Mag.* **11** 592
  Chandrasekhar S 1931b The maximum mass of ideal white dwarfs *Astrophys. J.* **74** 81
[3] Shapiro S L and Teukolsky S A 1983 *Black Holes, White Dwarfs and Neutron Stars (The Physics of Compact Objects)* (New York: Wiley)
[4] Landau L D 1932 On the theory of stars *Phys. Z.* **1** 285
[5] Baade W and Zwicky F 1934 Supernovae and cosmic rays *Phys. Rev.* **45** 138
[6] Oppenheimer J R and Volkoff G 1939 On massive neutron cores *Phys. Rev.* **55** 374
[7] Herold H and Neugebauer G 1992 Relativistic gravity research *Gravitational Fields of Rapidly Rotating Neutron Stars: Numerical Results (Lecture Notes in Physics 410)* ed J Ehlers and G Schäfer (Berlin: Springer) p 319
[8] Zimmermann H and Weigert A 1995 *ABC-Lexikon Astronomie* (Heidelberg: Spektrum)
[9] Stephani H 1977 *Allgemeine Relativitätstheorie* (Berlin: Deutsche) pp 213–15
[10] Soffel M H and Ruder H 1991 Der Doppel-Pulsar 1913+16 *Phys. Zeit* **1** 29
[11] Oppenheimer J R and Snyder H 1939 On continued gravitational contraction *Phys. Rev.* **56** 455
[12] Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (New York: Freeman)
[13] Ehlers J 1961 Beiträge zur relativistischen mechanik kontinuierlicher medien *Akad. Wiss. Lit. Mainz–Abh. Math.-Nat. Kl.* **11** 793
[14] Tolman R C 1934 Effect of inhomogeneity on cosmological models *Proc. Natl Acad. Sci., USA* **20** 169
[15] Datt B 1938 Über eine klasse von lösungen der gravitationsgleichungen der relativität *Z. Phys.* **108** 314
[16] Misner C W 1969 Gravitational collapse *Chrétien, Deser, and Goldstein* vol 1
[17] Hawking S W and Ellis G F R 1973 *The Large Scale Structure of Space-Time* (Cambridge: Cambridge University Press)
[18] Penrose R 1965 Gravitational collapse and space-time singularities *Phys. Rev. Lett.* **14** 57
[19] Hawking S W and Penrose R 1969 The singularities of gravitational collapse and cosmology *Proc. R. Soc. London* A **314** 529
[20] Neugebauer G and Meinel R 1995 General relativistic gravitational field of a rigidly rotating disk of dust: solution in terms of ultraelliptic functions *Phys. Rev. Lett.* **75** 3046

[21] Bardeen J M and Wagoner R V 1969 Uniformly rotating disks in general relativity
        Astrophys. J. **158** L 65
[22] Bardeen J M and Wagoner R V 1971 Relativistic disks I. Uniform rotation *Astrophys.
        J.* **167** 359
[23] Neugebauer G, Kleinwächter A, and Meinel R 1996 Relativistically rotating dust
        Helv. Phys. Acta **69** 472

Chapter 4

The environment of the Galaxy's central black hole

Mark R Morris
University of California, Los Angeles, USA

Back to the Milky Way, we now approach the Galactic Center black hole and study the innermost 100 parsecs (about 300 light years) around it. Various components interact with each other and may influence the feeding and activity of the black hole: stars, molecular and hot gas, magnetic fields, and supernovae. Star formation and black hole activity may be related in an ongoing cyclic behavior.

4.1 Introduction

At first glance, the Galactic Center region is apparently quite complex, especially when one's view spans a multitude of wavelengths. When the predominant structures are sorted out, however, it becomes evident that there is really more order there, in terms of scale, placement, and dynamics, than in almost any other place in the Galaxy. The structural elements that compose this region include:

(1) a central few-hundred-parsec zone of dense molecular clouds embedded in a hot, rarefied intercloud medium, all of which appears to be permeated by a strong magnetic field;

(2) massive short-lived star clusters, and clear signs of energetic bursts of current star formation;

(3) a sedate but dense cluster, or central cusp, of intermediate-age stars centered on the nucleus;

(4) several supernova remnants which are undoubtedly important for the dynamics of the region;

(5) a tumultuous, continuous disk of gas orbiting the nucleus on scales of a few parsecs; and

95

(6) of course, the centerpiece: the supermassive black hole lying at the dynamical center of the whole system.

The astrophysical phenomenology of the Galactic Center serves as a standard for understanding the nuclei of moderately gas-rich spiral galaxies in general and active galactic nuclei in particular. Of course, we have the advantage in the case of our own galaxy—located ~100 times closer than the next nearest nucleus— that its proximity allows us to observe important details that cannot be seen elsewhere with current technology. Among other things, these details include the structure of strong magnetic fields, the effects of strong tidal forces, the results of dynamical friction acting on stars, star clusters and clouds, and the extreme radiative environments near massive young clusters. A characterization of such phenomena is necessary to construct a reasonably complete picture of how the central black hole interacts with its environment. This chapter will describe the essential elements of the Galactic Center region, and what is now known about how each of them is relevant for the evolution and the activity of the central black hole.

4.2   The nuclear stellar bulge

The first near-infrared (NIR) mapping observation of the Galactic Center by Becklin and Neugebauer in 1968 revealed a central cluster of unresolved stars distributed with radius r as $r^{-1.8}$ on a scale of a few hundred parsecs (100 pc = 40 arcmin at a distance of 8 kpc). The infrared light from this cluster is dominated by emission from red giant stars. This central cusp of stars, also called the 'r^{-2} cluster' or the 'Nuclear Bulge', is superimposed on the much larger-scale (~2.5 kpc) bulge of our Galaxy. The Galactic Bulge has a bar morphology, as described in chapter 2. The three-dimensional morphology of the Nuclear Bulge, which resides well within the old Galactic Bulge, has not been determined, because of the difficulty in sorting out the highly variable extinction across the inner few degrees of the Galaxy.

Figure 4.1 shows a NIR image of the central few hundred parsecs of the Galaxy. The central cluster, or Nuclear Bulge, is evident among various patches of foreground extinction. With 30 magnitudes of visual extinction to the Galactic Center (a factor of 10^{12}), the extinction at 2 μm is about three magnitudes (a factor of ~10). This cluster reaches a maximum density exceeding 10^7 M$_\odot$ pc^{-3} in an uncertain core radius of ~0.06 pc (Alexander 1999; Genzel *et al* 2000).

The history of the Nuclear Bulge is likely related to that of the Central Molecular Zone (CMZ, described later and in chapter 2), given that they have comparable physical scales and that the current star formation rate in the CMZ is sufficient to populate the Nuclear Bulge, or r^{-2} cluster, over the lifetime of the Galaxy (Serabyn and Morris 1996). According to this hypothesis, the Nuclear Bulge is built up over a Hubble time by sustained star formation in the CMZ. The extent of the Nuclear Bulge is evident from the COBE images, shown in figure 4.2
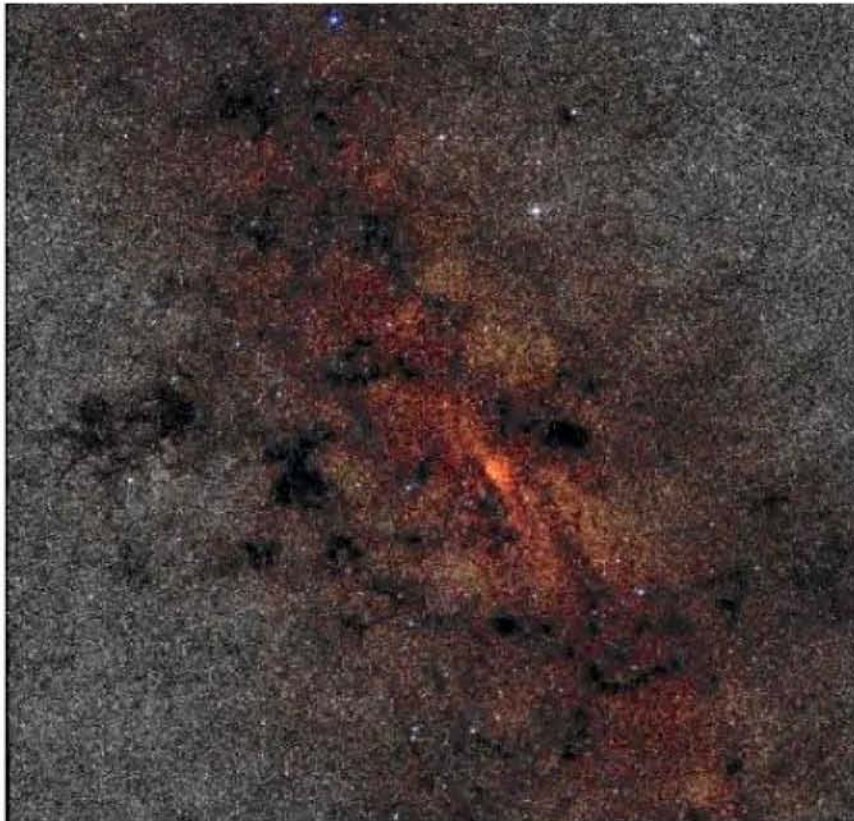
Figure 4.1. Three-color, NIR image of the central 330 pc of the Galaxy, from the 2MASS Two Micron All-Sky Survey (2MASS). The Galactic plane runs through this image at a position angle of about 30°. (2MASS is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.) See also color section.

(from Launhardt *et al* 1992). The fact that the Nuclear Bulge is substantially thicker than the cloud layer in the CMZ can be accounted for by the vertical diffusion of stars formed in the CMZ (Kim and Morris 2001). On timescales of a few Gyr, the scattering of stars by molecular clouds in the CMZ raises the stellar scale height to values comparable to those of the older OH/IR stars, about 44 pc (Lindqvist *et al* 1992).

The stars within the central parsec, where the Nuclear Bulge peaks, are shown in figure 4.3. This cluster, with an overall mass estimated at $\sim 10^9\ M_\odot$, dominates the gravitational potential outside of about 1 pc. The black hole's
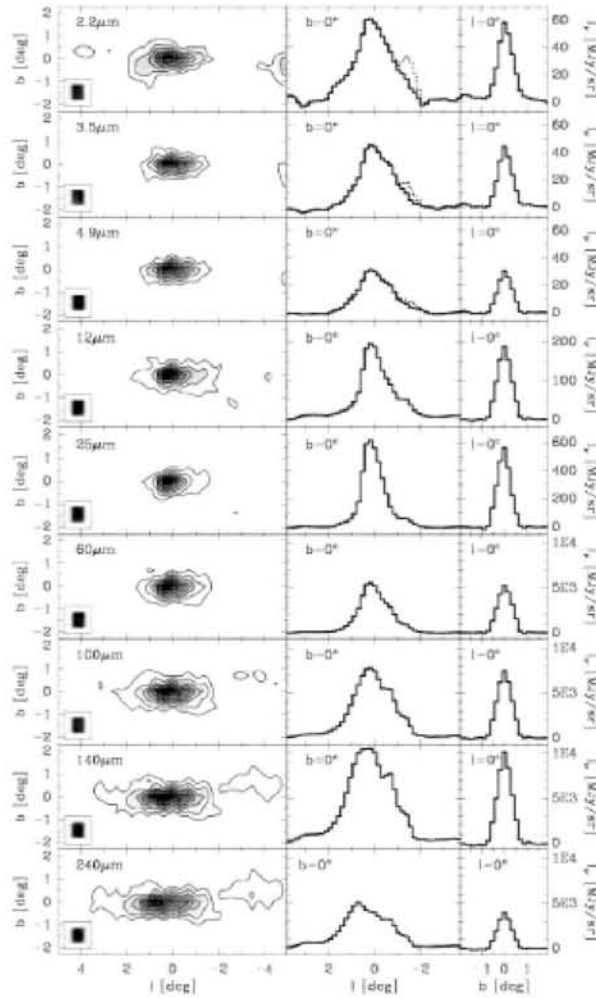
Figure 4.2. Surface brightness maps of the Nuclear Bulge, at nine wavelengths between 2.2 and 240 μm, as measured by the COBE satellite (from Launhardt *et al* 2002). These images are corrected for both emission and foreground extinction by dust in the Galactic Disk, and the emission from the much more extended Galactic Bulge has been subtracted. Lowest contour levels are at 10% of the maximum in the NIR maps and 5% in all other maps. Small boxes in the lower left of the maps show the DIRBE beam (HPBW = 0.7°) at the corresponding wavelength bands. The middle and right-hand panels show the corresponding longitude and latitude profiles at $b = 0°$ and $l = 0°$, respectively.

domain of influence is therefore quite small. In addition to the intermediate-age stars residing in the core of the Nuclear Bulge, the central parsec contains

Figure 4.3. Two-color (H: 1.65 μm and K$'$: 2.1 μm) image of the center of the Nuclear Bulge, showing a region 18 arcsc on a side (0.75 pc for a Galactic Center distance of 8 kpc). These data were acquired with the QUIRC camera behind the Hokupa'a adaptive optics system as part of the Demonstration Science Program of the Gemini North Telescope. See also color section.

a luminous young cluster dominated by emission-line stars (Krabbe *et al* 1995; Paumard *et al* 2001). The difficulty of estimating the core radius of the cluster is attributable to the high luminosity of the emission-line stars, which overwhelms the light from the older stars.

A very compact cluster of moderately luminous stars ($m_K = 14$–16) lies within 0.5$''$ of the black hole: the Sgr A* (IR) cluster, discussed in detail in chapter 8. The blue color and featureless K-band spectra of these objects (Genzel *et al* 1997; Gezari *et al* 2002) indicate that they may be massive, young stars, although if they are lower-mass stars heated by some process peculiar to the immediate vicinity of the black hole, then they may represent the innermost members of the Nuclear Bulge.
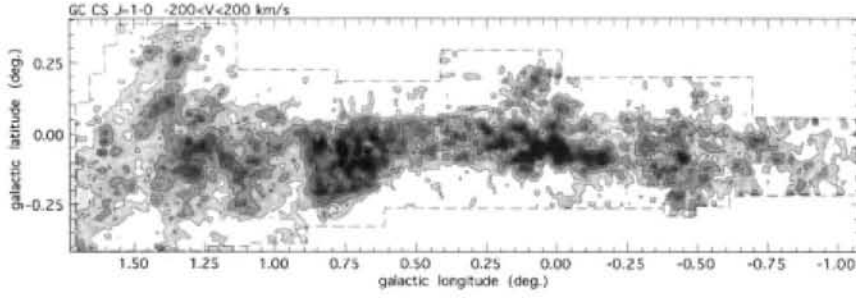
Figure 4.4. Map of integrated CS $J = 1$–0 line emission, showing the spatial extent of relatively dense clouds in the CMZ, as measured by the 45-m telescope of the Nobeyama Radio Observatory (from Tsuboi *et al* 1999).

4.3   The Central Molecular Zone

Various lines of evidence indicate that a total of $5 \times 10^7 \, M_\odot$ of relatively dense molecular gas are present in the CMZ, lying predominantly within ± 150 pc of the Galactic Center. This represents a much stronger concentration of interstellar matter than anywhere else in the Galaxy. Figure 4.4 shows the bulk of the CMZ as revealed by CS emission, which is a probe of moderately dense clouds ($n_{H_2} \sim 10^4 \, cm^{-3}$). The CMZ has already been described by Hüttemeister (chapter 2); here we make a few additional points:

1. *Tidal shear.* The clouds in the CMZ are easily sheared into tidal streams by the relatively strong differential gravitational forces present in the central few hundred parsecs. A likely example is the 'Galactic Center Bow', shown in figure 4.5, from Tsuboi *et al* (1999). This apparently single molecular structure appears to extend over 300 pc in projection, and probably undergoes most of a complete turn around the Galactic Center. The stability of clouds against such tidal shear depends on their density. Güsten (1989) derived the following density criterion for cloud stability as a function of galactocentric radius, R:

$$n \geq 10^4 \, cm^{-3} (75 \, pc/R)^{1.8}. \tag{4.1}$$

This condition helps explain why the density of molecular clouds in the CMZ is typically substantially larger than the average molecular cloud density in the Galactic disk. Only the densest molecular clouds in the CMZ, such as Sgr B2, are immune to shear, and thus well localized, and even in those cases, the dense cloud cores are surrounded by a sheared halo.

2. *Residence time.* Clouds in the CMZ are transient, because their orbital angular momentum is lost by dynamical friction on timescales of several times 10^8 years (Stark *et al* 1991). Magnetic torques can cause angular momentum loss on a comparable timescale, given the exceptionally strong magnetic field in the CMZ (Morris 1994, and later). Consequently, the CMZ must be constantly
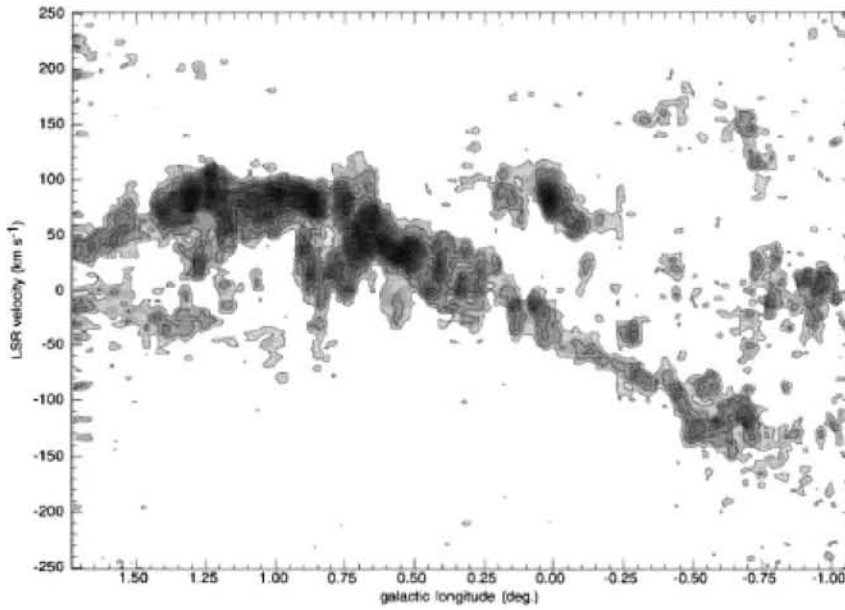
Figure 4.5. CS $J = 1$–0 position-velocity diagram following the ridge of the 'Galactic Center bow' (Tsuboi *et al* 1999). The ridge is defined by the points $[l, b] = [-1.07, -0.20]$, $[-0.75, -0.03]$, $[-0.37, 0.05]$, $[0.50, 0.05]$, and $[1.73, -0.17]$. Note how this structure is continuous over $2.5°$ (375 pc) and 250 km s^{-1}, suggesting that this is a single structure which has been subjected to tidal shearing. It probably wraps at least halfway around the Galactic Center.

replenished, presumably by gas migrating inward from the Galactic disk (Morris and Serabyn 1996; Morris 2001).

What, then, is the fate of the molecular gas which moves inward through the CMZ? Sitting at the bottom of the Galactic potential well, it has only a few possibilities: star formation, ejection in a hot galactic wind, and accretion onto the central black hole. The latter possibility cannot account for any but a tiny fraction of the inflowing gas. The mass of the central black hole and the currently inferred accretion rate onto the black hole are too small by several orders of magnitude for the black hole to be a significant sink for CMZ gas. A thermal galactic wind is a possible contributor (discussed later), but the dominant sink for the CMZ gas is probably star formation, occurring at a rate of a few tenths of a solar mass per year.

3. *Asymmetry*. The CMZ is quite asymmetric about the Galactic Center, with most of the gas, perhaps 70% of it, lying at positive Galactic longitudes. The thermal dust emission which follows the molecular gas illustrates this asymmetry, as can be seen in the longer-wavelength images of figure 4.2. This asymmetry,

also present in the velocity field, could be the result of an $m = 1$ sloshing mode in the gas of the CMZ (discussed by Morris and Serabyn 1996). Notably, the stars do *not* appear to show a comparable asymmetry, although it would be interesting to seek an asymmetry in the stellar distribution on the scale of the Nuclear Bulge by carefully accounting for extinction. If an $m = 1$ oscillation is present in the stars and gas at the Galactic Center, corresponding to coupled orbits of the CMZ and Nuclear Bulge about a common center of mass, then the stars and the central black hole would participate in this oscillation, which might manifest itself in terms of their velocities.

4.4 Hot gas

Extended X-ray emission from the central 300×200 pc of the Galaxy reveals that the CMZ is embedded in a corona of hot, low-density gas that probably occupies most of the volume of the central zone, and is roughly coextensive with the CMZ (Kawai *et al* 1988; Koyama *et al* 1989, 1996; Yamauchi *et al* 1990; Sunyaev *et al* 1993; Markevitch *et al* 1993). The hot gas is evidenced by both the X-ray continuum and the 6.7 keV line emission from helium-like iron (24 times ionized). A recent mosaic of the continuum emission made with the Chandra X-Ray Observatory is shown in figure 4.6. It illustrates the mix of diffuse emission and large numbers of point sources, most of which are X-ray binary stars. Until recently, it was thought that the gas temperature in the X-ray emitting gas was $\sim 10^8$ K because of the strength of the 6.7 keV line, but that interpretation has recently been called into question by Wang *et al* (2002). With the Chandra X-Ray Observatory, they have mapped a $2°$ by $0.8°$ region about the Galactic Center with $1''$ spatial resolution, and have found that much of the 6.7 keV line emission emanates from discrete stellar sources (presumably close binary systems), rather than from the diffuse structures. Wang *et al* argue that the temperature of the hot gas is therefore more like 10^7 K, a much less demanding constraint in terms of the requisite heating source. One of the most important consequences of lowering the temperature to 10^7 K is that the gas is bound to the Galaxy. At 10^8 K, the gas would be unbound, implying a thermal, galactic wind. In this case, the galactic wind could well be a major sink for the gas migrating inwards through the CMZ. The XMM satellite observatory should soon provide images of the full extent of the coronal gas bubble.

4.5 The Galactic Center magnetosphere

The magnetic field in the inner few hundreds of parsecs of the Galaxy has been somewhat of a surprise, inasmuch as it has been found to be unexpectedly strong (milligauss versus a few microgauss in the Galactic disk), apparently highly ordered, and characterized by a completely different geometry (dipolar) from the field in the Galactic disk (toroidal).

Figure 4.6. False-color mosaic image of X-ray emission from the Galactic Center region, from Wang *et al* (2002). This image covers a band of about $2° \times 0.8°$ in Galactic coordinates (300×120 pc), and is centered at $l = -0.1°$, $b = 0°$, roughly the location of the Sgr A complex. The three energy bands are 1–3 keV (shown in red), 3–5 keV (green), and 5–8 keV (blue). The image was compiled from 30 separate pointings acquired with the Advanced CCD Imaging Spectrometer (ACIS-I). The intensity is plotted logarithmically to emphasize low-surface-brightness features.

There are several ways of probing the magnetic field at the Galactic Center, and all have been used to study the Galactic Center field (Morris 1994, 1998; Morris and Serabyn 1996 and references therein):

(1) the Zeeman effect in the spectral lines of OH and HI,
(2) Faraday rotation of polarized background radio emission,
(3) the orientation of the polarization vectors of the thermal emission from magnetically aligned dust grains,
(4) the filamentary morphology of synchrotron-emitting structures and
(5) the intrinsic orientation of the polarization vectors in the synchrotron-emitting structures.

An example of an application of the third method is shown in figure 4.7, taken from Aitken *et al* (1998).

Filamentary radio structures which delineate the magnetic field (method 4) are illustrated in figure 4.8. These structures are representative of about a dozen filamentary radio systems, the most prominent of which is the bundle of filaments constituting the Galactic Center Radio Arc (situated at the top of figure 4.8, and not well represented in this radiograph because of its distance from the center of the field; see Yusef-Zadeh *et al* 1984). These non-thermal radio filaments (NTFs) generally share several characteristics (Morris 1996):

(1) Their radio emission is highly polarized synchrotron radiation, and when the polarization vectors are corrected for Faraday rotation by the intervening
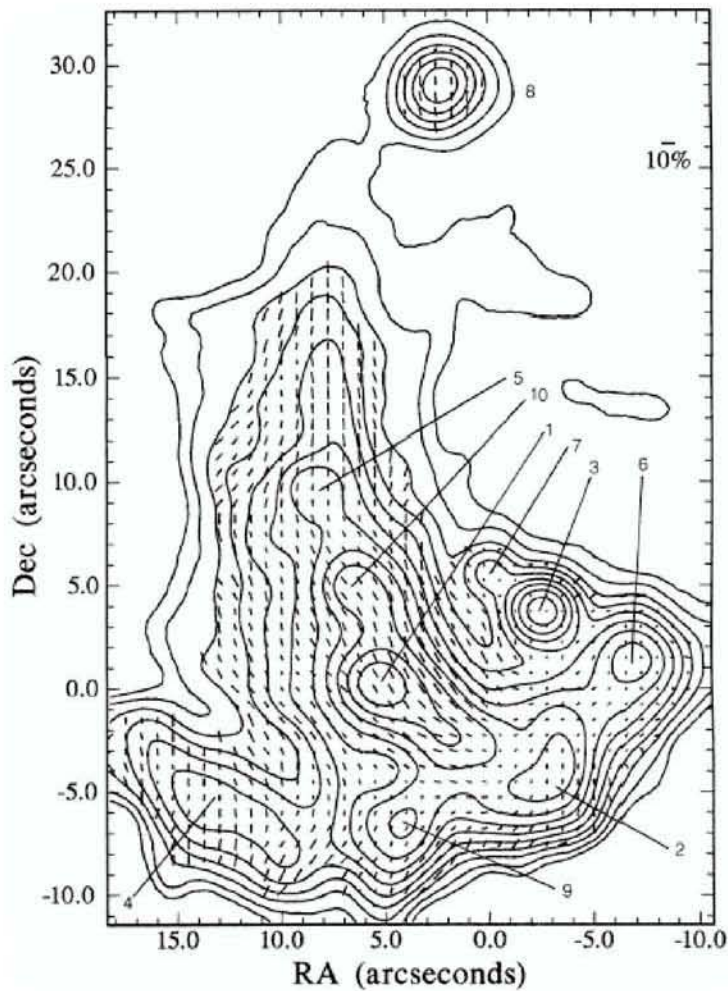
Figure 4.7. Vectors orthogonal to the 12.5 μm polarization vectors measured toward Sgr A West by Aitken *et al* (1998), superimposed on contours of the surface brightness at that wavelength. These vectors are probes of the magnetic field direction to the extent that the emitting dust grains have been collectively aligned by the magnetic field. Note how the magnetic field follows the Northern Arm, the north–south structure which is a stream of dust and gas falling towards, and orbiting partially around, the black hole (see figure 4.10).

medium, they are found to be perpendicular to the filaments, indicating that, not surprisingly, the magnetic field is aligned with the filaments.

(2) They are typically quite long: 20–50 pc, and narrow: less than 0.3 pc.

(3) With one exception, the NTFs all have a gentle, smooth curvature with no
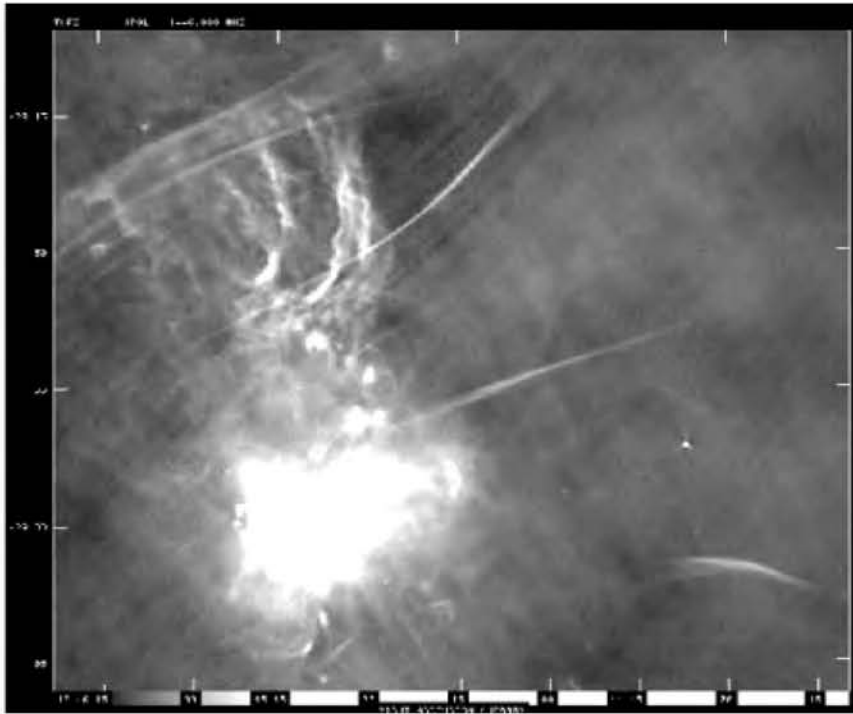
Figure 4.8. The radio 'threads' observed toward the Galactic Center with the Very Large Array (VLA) at a wavelength of 20 cm (Lang *et al* 1999). Other filamentary systems are evident as well. Note the strikingly large scale of the radio filaments; the 5 arcmin between tick marks on the vertical axes corresponds to 12.5 pc.

gross deformations. The exception is the 'Snake', or G359.1-00.2 (Gray *et al* 1995), which shows a few kinks along its length.

(4) All of the NTFs within a projected distance of 150 pc of the Galactic Center are approximately perpendicular to the galactic plane.

The generally smooth curvature of the NTFs has been used as evidence for a high rigidity of the magnetic field (Yusef-Zadeh and Morris 1987a). That is, the magnetic field strength must be of the order of a milligauss in order to not show distortions resulting from inevitable (and observed) interaction with the tumultuous interstellar medium at the Galactic Center. A milligauss magnetic field has a tremendous pressure compared to other sources of interstellar pressure in the Galactic Center region, so it is unlikely that the NTFs are isolated, because in that case, their internal pressure would cause them to expand explosively. Self-confinement of isolated, force-free magnetic field configurations has been considered, but this hypothesis does not solve the problem that the filamentary

structures will expand and dissipate without some confining medium. The confinement of the field is assured, however, if the magnetic field in the central 150 pc of the Galaxy is ubiquitous and approximately uniform. Then the NTFs represent locations where the more-or-less uniform magnetic field is 'illuminated' by the local injection of relativistic electrons. The filamentary morphology can be understood in terms of the fact that the large-scale field constrains these electrons to diffuse only parallel to the magnetic field lines. The overall picture indicated by the orientation of all the filaments is that of a dipole field with perhaps some divergence above the Galactic plane. The confinement and stability of a strong central field have been discussed recently by Chandran (2001).

The NTFs define the magnetic field in the intercloud medium. Within molecular clouds, the magnetic field orientation has been measured using the far-infrared polarization of the thermal emission from dust grains and found to be quite different: largely parallel to the Galactic plane (Hildebrand *et al* 1990, 1993; Morris *et al* 1992; Morris and Serabyn 1996; Novak *et al* 2000). While the magnetic field strength is not directly determined by this method, the striking uniformity of the orientation of the polarization vectors measured at adjacent positions suggests that it is at least as strong as the field in the intercloud medium (following the arguments of Chandrasekhar and Fermi 1953). The orientation of the field within clouds can be attributed to the shear to which molecular clouds are subjected there (Aitken *et al* 1998; Morris *et al* 1992).

The rough orthogonality of the cloud and intercloud fields raises the possibility that field line reconnection could be a mechanism for particle acceleration within the filaments. This mechanism was explored by Serabyn and Morris (1994), who argued that the reconnection process is most effective where the cloud surface is ionized, presumably by a fortuitously nearby star, and where the cloud is moving with respect to the ambient magnetic field. The HII region at the cloud surface provides two things: turbulence, which mixes the cloud and intercloud fields and fosters their reconnection; and the electrons which are thereby accelerated. So far, every NTF which has been sufficiently well studied can be connected with a surficially ionized molecular cloud somewhere along its length (e.g. Uchida *et al* 1996; Kramer *et al* 1998; Staguhn *et al* 1998). However, further evidence is needed to establish this hypothesis, such as a demonstration that the spectral indices of the NTFs always steepen away from the presumed acceleration site as the emitting electrons lose energy.

How is the strong, vertical magnetic field at the Galactic Center produced? No dynamo models have yet successfully produced a strong dipole field at the Galactic Center. The one extant model that can account for it posits that the central field simply results from the slow inward migration, over a Hubble time, of partially ionized gas, to which the vertical component of the protogalactic magnetic field is effectively frozen (Sofue and Fujimoto 1987; Morris 1994; Chandran *et al* 2000). The field component parallel to the Galactic plane follows a different history; it can be lost by drifting vertically out of the thin Galactic disk by ambipolar diffusion, but along the way it can be amplified by differential

rotation and other dynamo processes. In any case, the vertical component is the only one which can be concentrated at the center.

All the evidence we have from measurements made at large spatial scales indicates that matter that accretes onto the black hole from the interstellar medium is relatively strongly magnetized. Even if the accreting matter comes predominantly from the winds of the massive stars near the central black hole, those stars, and thus their winds, are presumably quite strongly magnetized, given that they recently formed out of such highly magnetized material as is present in the CMZ.

4.6   The circumnuclear disk and Sagittarius A West

Well within the central molecular zone lies a disk of molecular gas which surrounds and orbits the central black hole. This distinctive structure is important because of its potential for providing matter to the black hole. At present, this circumnuclear disk (CND) has an inner cavity with a radius of about 1 pc, and the disk can be followed out to as much as 7 pc in some directions with molecular line observations (Morris and Serabyn 1996). The CND is inclined by about 60° to the line of sight (see figure 4.9), so it is in fact tilted with respect to the Galactic plane (Güsten *et al* 1987; Marshall *et al* 1995). This is a very fortunate circumstance, because, if the CND were aligned with the Galactic plane, or if it were edge-on to our line of sight in some other plane, then the extinction through this disk would have made the study of stars and gas in the central parsec extremely difficult.

The CND was discovered in the far-infrared, at a wavelength of 100 μm (Becklin *et al* 1982), where, because of limb-brightening, it appears as two lobes of emission straddling Sgr A*—the compact radio core (see chapter 11). The prominent far-infrared emission from the CND emanates from the warm dust heated by the extremely luminous, hot stars in the central stellar cluster (Davidson *et al* 1992; Dent *et al* 1993; Telesco *et al* 1996). A layer of gas at the abrupt inner edge of the disk is ionized (at least where the UV starlight is able to reach that inner edge; the gaseous features comprising the H II region Sgr A West, described later, apparently block some of the ionizing radiation from reaching the CND). Immediately exterior to the ionized ring is a ring of shocked molecular hydrogen, as is shown in figure 4.9 (Yusef-Zadeh *et al* 2001). The CND is a clumpy structure on all scales which have been used to observe it. This implies, in principle, that UV radiation from the enclosed stars can penetrate much further into the disk than would be possible if the medium were uniform. The penetration of UV radiation also leads to a thick photodissociation region in the interior portions of the CND, where atomic and molecular gas coexist (C II, Lugten *et al* 1986; H_2, Gatley *et al* 1986, Yusef-Zadeh *et al* 2001; O I, Jackson *et al* 1993; C I, Serabyn *et al* 1994; HCN, Marshall *et al* 1995; other molecules, White 1996 and Wright *et al* 2001). The detailed microstructure of the CND has been modeled by Vollmer and Duschl (2000, 2001a, b).
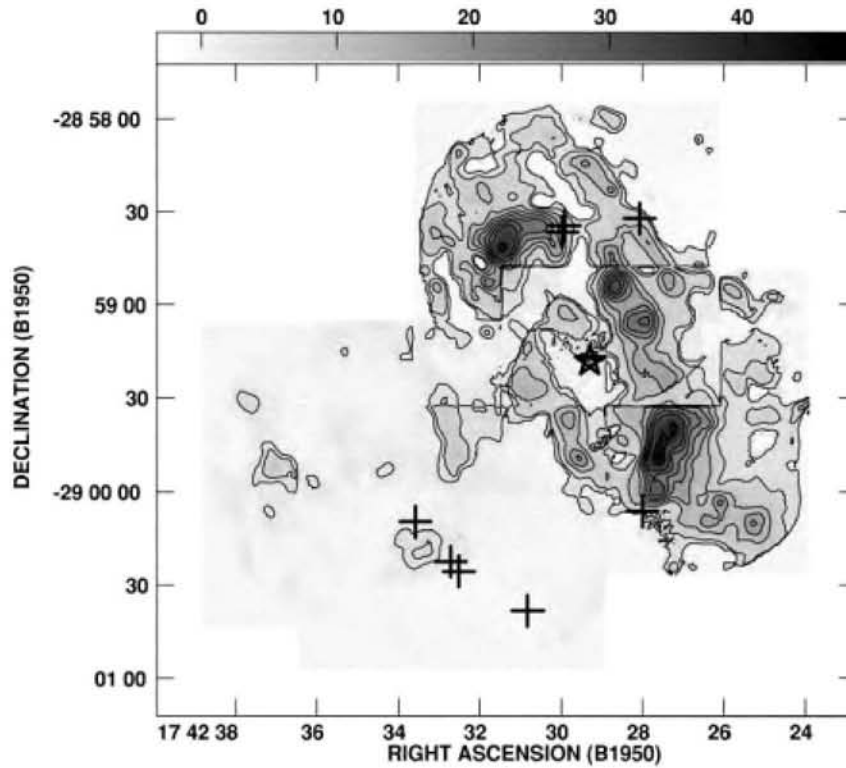
Figure 4.9. The circumnuclear disk, as seen in NIR emission from shocked molecular hydrogen by the NICMOS camera on the Hubble Space Telescope (from Yusef-Zadeh *et al* 2001). This mosaiced image shows the gray scale and contours of velocity-integrated H_2 S(1) 1–0 line emission. The star at the center of the configuration shows the location of the radio counterpart to the Galactic black hole, Sgr A*. The square in the ring is due to artifacts of mosaicing. The plus signs mark the positions of 1720 MHz OH masers, which Yusef-Zadeh *et al* (1996) argue are due to shocked molecular gas. The shocked molecular hydrogen arises primarily near the inside edge of the circumnuclear disk, where it is presumably shocked by the impact of the high-velocity winds from the hot stars residing in the central parsec (see text).

The CND has a relatively flat rotation curve, starting at about 110 km s^{-1} at its inner edge (Roberts and Goss 1993). The determination of the rotation curve requires knowledge of the possibly radius-dependent orientation of the disk plane (Marshall *et al* 1995), and this is not yet well determined. This orientation depends in detail on the recent history of the CND, for example, whether it is a result of slow accretion of material migrating inwards from the CMZ or whether it is a dispersion ring resulting from the tidal alteration of a dense cloud which

recently passed within a few parsecs of the central black hole and got sheared into a self-intersecting orbit (Sanders 1998). The CND is apparently approximately circular at its inner edge, but its outer contours deviate strongly from axisymmetry (e.g. Morris and Serabyn 1996). This can be accounted for by Sanders' model, but an alternative effect that merits consideration is that the blast wave of the supernova remnant Sgr A East (see later) has recently passed by the CND, and has deformed its originally more symmetric geometry (Maeda *et al* 2002).

CNDs are common to a large fraction of gas-rich spiral galaxies, notably many with active galactic nuclei. The Hubble Space Telescope (HST) has imaged many of them, and their physical scales tend to be larger (~100 pc) than that of our Galaxy's CND, but that is likely to be a selection effect because that is the scale corresponding to the resolving power of the HST. The importance of all these disks is that they are likely to serve as reservoirs for fuel for the central black holes of these galaxies. These disks are also likely to be rich in star formation, and may be especially important in starburst galaxies. Our Galaxy's CND is currently the only gaseous structure which might conceivably fuel star formation in the central parsec, although there is no evidence that the CND is forming stars at the present time.

Sagittarius A West consists of all the ionized gas at the inner edge of the CND and the plasma structures within the central cavity of the CND. In radio continuum images (figure 4.10) and images made from infrared lines characteristic of H II regions (including radio recombination lines, Brackett-γ, and [Ne II] 12.7 μm), Sgr A West assumes the form of a three-armed, triskelian, pattern, probably as the result of the superposition of several more-or-less independent structures projected along our line of sight, rather than because there is a real three-armed structure in this direction.

At least two of the features in Sgr A West—the Northern and Eastern[1] Arms—appear to be infalling streams of gas. Their kinematics have been modeled in terms of gas accelerating inwards toward the black hole on highly eccentric orbits (Serabyn *et al* 1988; Herbst *et al* 1993; Roberts *et al* 1996; Morris and Maillard 2000; Vollmer and Duschl 2000). The leading apex of these streams is well defined and indicates that, while close to the central black hole, the orbital motion has not yet carried the apex past periapse. One might anticipate that, on a timescale as short as 10^4 years, at least one of these streams will self-intersect and create a dispersion ring well within the inner cavity of the CND. If the two streams collide, the dynamics will be more complex, and the angular momentum loss will lead to an even more compact disk around the black hole. In either case, the accretion rate onto the black hole is likely to be considerably enhanced over its present, relatively small rate. The origin of these streams is still unclear; possibilities include:

(1) a magnetohydrodynamic instability occurring at the inside edge of the CND,

---

[1] Note that in astronomical maps and images East is on the left.

$$2''$$
$$0.08\text{pc}$$
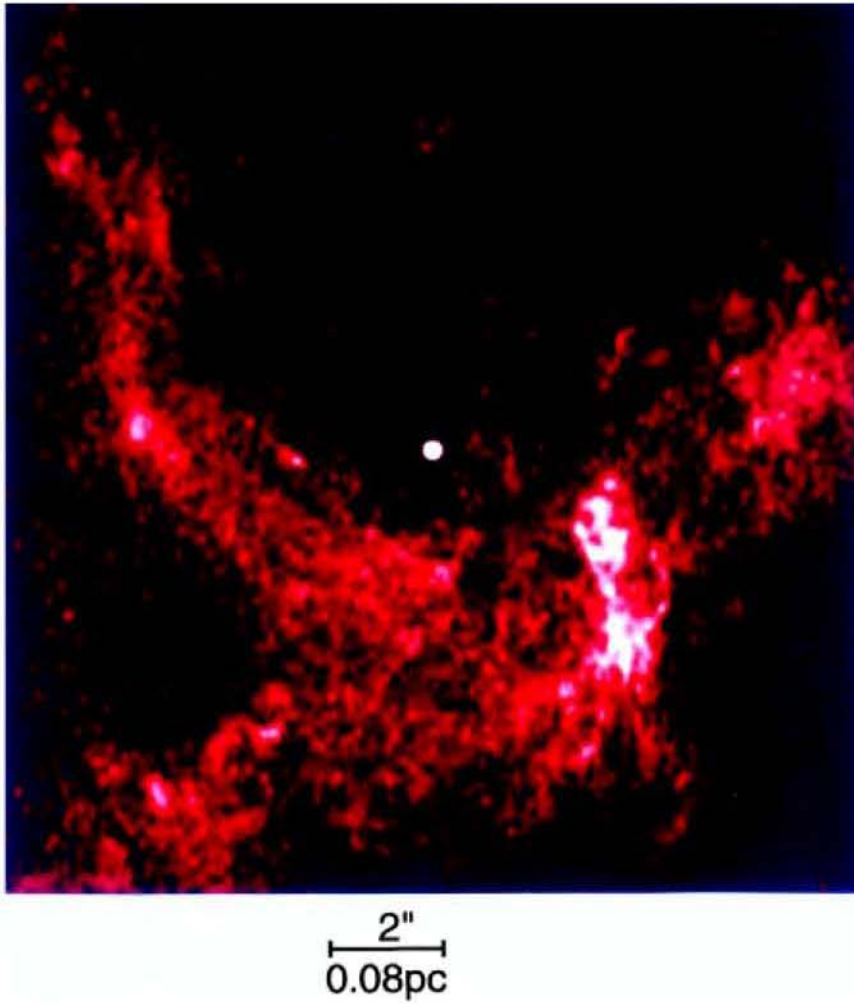
Figure 4.10. Radio continuum image of the central portion of Sgr A West, made at 1.3 cm wavelength with the VLA by Zhao and Goss (1998). The spatial resolution of this image is $0.1''$. Sgr A* is the prominent point source at the center of the image, while the Northern Arm falls from the northeast (top left) around Sgr A* to the south, where it joins the 'mini-cavity', centered $\sim 3''$ to the southwest of Sgr A*. The Eastern Arm comes in from the southeast, and the extended emission feature cutting through the mini-cavity from southeast to northwest is the 'bar'. See also color section.

(2) collision of a small cloud with the CND, leading to local angular momentum loss, with subsequent infall (Jackson *et al* 1993), and

(3) a low-angular-momentum gas cloud from well beyond the CND, now

approaching its pericenter distance. The presence of presumably infalling molecular gas streams on somewhat larger scales (e.g. McGary *et al* 2001) would favor the last two possibilities.

4.7   Star formation

Star formation in the CMZ seems to be rather different from that seen elsewhere in the Galaxy. For one thing, one finds in this region the most extreme young star clusters in the Galaxy. There are three remarkable, massive, short-lived star clusters known in this region: the Arches, the Quintuplet and the central parsec cluster. These clusters have ages $< 10^7$ years and masses on the order of 10^4 $M_\odot$. All three have a large number of unusually massive, windy stars, including a mix of Wolf–Rayet stars, Luminous Blue variables, Ofpe stars, and a sizeable population of OB stars (Krabbe *et al* 1995; Najarro *et al* 1997; Figer *et al* 1999a, b; Paumard *et al* 2001). Other exotic categories of very luminous stars are also present, all falling under the rubric of helium emission-line stars, after the presence of the 2.06 μm He emission line, an indicator of a substantial, strong wind. The precise mix of the more evolved stars—the WR stars and LBVs—is an indicator of the cluster age.

The Arches cluster is the youngest and most extreme (figure 4.11). It has $\sim$160 O stars, a total luminosity exceeding 10^8 $L_\odot$, and a Lyman continuum production rate of $\sim$4 $\times$ 10^{51} s^{-1} (Figer *et al* 1999b). With this radiation field, the Arches cluster dominates the local heating and ionization of the interstellar medium, and in fact the region surrounding the Arches cluster is the most luminous portion of the Galactic Center region at mid-infrared wavelengths (e.g. Shipman *et al* 1997), because much of the luminosity of the Arches is reradiated at mid-infrared wavelengths. Furthermore, in the radio regime, the unusually large HII region known as the arched filaments—linear, ionized features lying at the surface of a molecular cloud (Lang *et al* 2001, 2002)—is apparently attributable to the Arches cluster.

The youth of these clusters is assured, because, by virtue of being within about 40 pc of the Galactic Center, they are subject to tidally induced evaporation on timescales not much larger than their ages, $\sim$10^7 years (Kim *et al* 1999, 2000; Portegies-Zwart *et al* 2002). The disintegration of these clusters is hastened by stellar evolution; the large rate of mass loss by the massive stars steadily reduces the cluster mass, which in turn reduces the tidal radius.

The cluster of young stars within the central parsec is not subject to tidal disruption because it is not a bound system. It consists of a grouping of massive stars having independent, phase-mixed orbits in the potential well created by the black hole and the central cluster of nuclear bulge stars. The youth of the emission-line stars in the central cluster raises a troubling question about where they were formed. It ordinarily takes a time far longer than the age of these stars to bring them individually into the central parsec from larger distances by
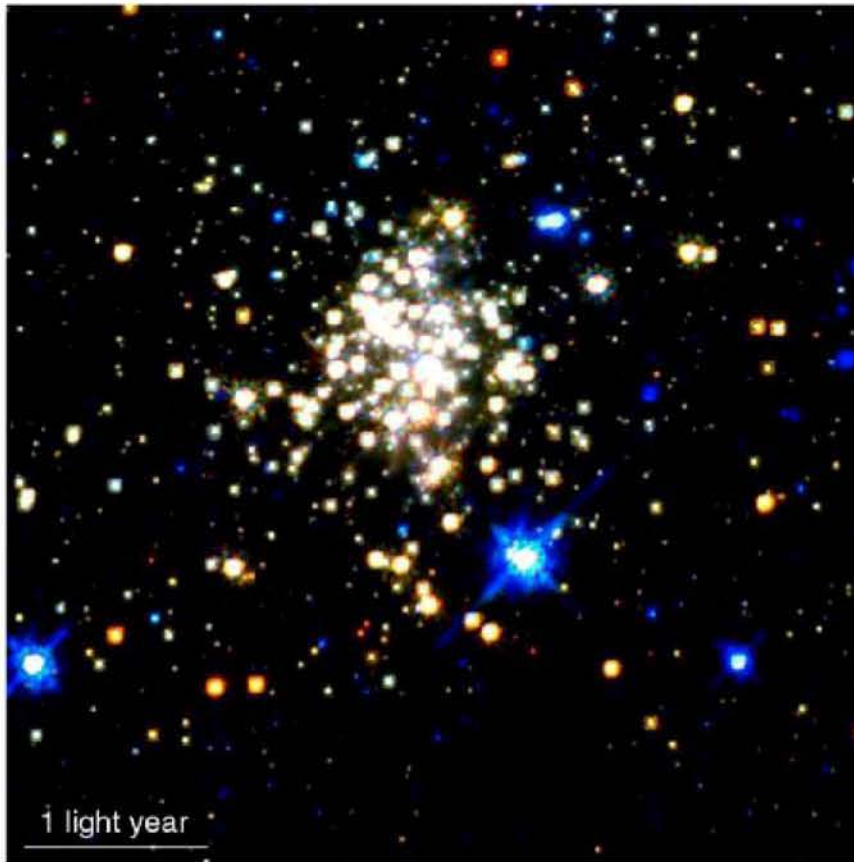
Figure 4.11. The Arches cluster, observed with HST/NICMOS (Figer *et al* 1999b). This false-color image was made by combining images made with three near-infrared filters. See also color section.

relaxation processes such as dynamical friction, or, equivalently, mass segregation (Morris 1993). However, *in situ* star formation is problematical because of the strong tidal forces exerted by the central black hole. At a distance of 0.25 pc, the typical distance of the luminous, He emission-line stars from Sgr A*, the limiting Roche density is $\sim 10^{10}$ H atoms cm^{-3}. This is 10^4–10^5 times denser than any gas presently observed near the Galactic Center, and there are serious problems with understanding how gas could be compressed to such high densities in such a warm, turbulent region, except possibly by the sudden release of an enormous quantity of mechanical energy, presumably by a dramatic accretion event onto the black hole. This scenario was considered by Morris *et al* (1999), who hypothesized a limit cycle of activity within the central parsec. Assuming

that the CND is a long-lived configuration continuously fed from the outside, and noting that the natural evolution of the inner edge of the turbulent, magnetized CND is to move inward as a result of viscous evolution, Morris *et al* suggested that the collective winds from all the young stars in the central parsec exert a ram pressure on the inside edge of the disk which is sufficient to impede that inward migration (except perhaps for Rayleigh–Taylor instabilities such as these referred to earlier in the discussion of the arms of Sgr A West). However, when the most massive young stars in the central cluster finish their evolution, on timescales of $\sim 10^7$ years, the winds will die out and the inner edge of the CND will proceed inward. Eventually, the CND will converge upon the central black hole. If this leads to a sudden increase in the black hole's accretion rate, then the release of accretion energy will be explosive, and the portions of the CND near the central black hole will be strongly compressed. Whether that compression is sufficient to overcome the tidal forces and allow self-gravity of the compressed layer to form stars remains to be seen, but if it does happen like that, then the newly-induced generation of stars and their stellar winds will establish a new dynamical equilibrium with the inside edge of the CND, initially joining with the outpouring of accretion energy from the black hole to evacuate the center of the CND. Thus the cycle would start anew, with a quasi-static equilibrium again resembling the current situation in the central parsec.

As an alternative, Gerhard (2001) recently explored the hypothesis that the young stars in the central cluster formed as part of a massive, Arches-type cluster originally located a few tens of parsecs away from the center. Because the timescale for spiraling inward to the central parsec as a result of dynamical friction is inversely proportional to the mass of the cluster, the cluster will move into the central parsec on a sufficiently short timescale if it is massive enough. Gerhard found that a mass as large as 10^6 $M_\odot$ is needed to account for the central young cluster (plus its parent cloud, if that cloud accompanies it most of the way into the center) if it starts as far out as the Arches cluster. While this mass exceeds that of the Arches cluster by two orders of magnitude, it is not unprecedented: super star clusters evidenced in starburst galaxies of various kinds have masses of 10^5–10^6 $M_\odot$ (e.g. Ho and Filippenko 1996; O'Connell *et al* 1994, 1995; Tacconi-Garman *et al* 1996; Turner *et al* 2000; Maoz *et al* 2001). However, in the Galactic Center, there is no evidence yet for the stellar tidal debris of young stars that would be left behind at radii beyond a parsec as the massive cluster migrated inward. This point is emphasized by Kim and Morris (2002), who use an N-body code to model the dynamics of a massive cluster at the Galactic Center. They confirm that it is possible to bring the remnant core of a cluster into the central parsec of the Galaxy if the cluster starts out massive enough, but, in general, the process should leave a halo of tidally stripped young stars throughout the inner several parsecs. Currently, there is no evidence for a population of young stars beyond the inner parsec.

The three remarkable clusters in the Galactic Center region, and the presence of super star clusters in the nuclear regions of starburst galaxies, suggest a

particular mode of star formation that differs from that usually found in the Galactic disk (Morris 2001). Unless we live at a peculiar time, we must imagine that compact massive clusters like the Arches and Quintuplet form often and represent an important channel for populating the nuclear bulge. If other such clusters are currently forming, their formation sites have not yet been identified, although the star-forming core of the molecular cloud Sgr B2 is currently forming a fairly massive star cluster which may qualify for being related to the existing clusters (Mehringer *et al* 1993; Gaume *et al* 1995). The compactness and high mass of the Arches cluster or any other starburst cluster raises the issue of the timescale over which the cluster must have formed. The violence implied by the formation of hundreds of O stars within a few tenths of a parsec, including protostellar jets and winds and ionized gas flows at ionization fronts, is likely to quickly shut off star formation once the process begins. Indeed, these clusters may begin formation on the scale of a Jeans mass ($\sim 10^5$ years), and then fragment hierarchically to stellar masses on a free-fall timescale. If so, then there is little room for the persistence of straggler gas clumps; the released gravitational energy rushing outward from the star formation cataclysm will commit any gas clump to immediate collapse or to a quick oblivion via ionization and Kelvin–Helmholtz instabilities. Massive starburst clusters must be quite close to coevality, so their use as probes of the IMF (initial mass function, i.e. the distribution of stellar masses when they form) should be little affected by a spread in stellar ages.

The core collapse of compact, massive, young clusters could provide a means of producing intermediate-mass black holes via stellar coalescence. Over time, dynamical friction would bring such black holes into proximity and eventual coalescence with the central black hole, providing a means of growth for black holes in Galactic nuclei (Ebisuzaki *et al* 2001).

Of course, massive star clusters are not the only way that stars form in the Galactic Center region, and probably not even the dominant way. Many individual compact HII regions and emission line stars have been identified and studied there (Morris 1993; Figer *et al* 1994; Liszt and Spiker 1985: Lis *et al* 1994; Zhao *et al* 1993), although no generalizations about the properties of the stars have yet been offered.

4.8   A provocative supernova remnant: Sgr A East

One of the important activities at the Galactic Center is that represented by the apparent supernova remnant, Sagittarius A East, which is relatively close to the central black hole. This non-thermal shell source has been well studied at radio wavelengths for some time (Ekers *et al* 1983; Yusef-Zadeh and Morris 1987b; Mezger *et al* 1989; Pedlar *et al* 1989). The radio to submillimeter studies reveal that Sgr A East lies largely behind Sgr A West, although perhaps not entirely; some portion of it must lie in front. In addition, the shell source surrounds Sgr A* and Sgr A W in projection, though its center is displaced from Sgr A* by about

2.5 pc. The likelihood that Sgr A East has affected the dynamics and the geometry of the CND has been discussed by Morris and Serabyn (1996) and by Yusef-Zadeh *et al* (1999).

The idea that Sgr A East is a supernova remnant has been debated in the literature because, while it has the appropriate morphology, its energy requirements are unusually large (e.g. Mezger *et al* 1989). This led some to consider that Sgr A East might have resulted from the release of energy from the central black hole, in spite of the fact that it is offset from Sgr A* by several parsecs. One hypothesis that accounts for that offset posits that Sgr A East resulted from the tidal disruption of a star by the central black hole (Khokhlov and Melia 1996). Such an event should occur every 10^4 to 10^5 years, consistent with the 10^4-year expansion age of the Sgr A East shell. When this happens, about half of the stellar mass is ejected as a spray into a large solid angle with an energy well in excess of that of a supernova. The subsequent expansion of the relativistic ejecta could thereby produce a shell source resembling Sgr A East. More recently, Sgr A East was studied in detail by Maeda *et al* (2002) with the Chandra X-Ray Observatory (CXO). The X-ray image (figure 4.12), when combined with the radio data, suggests that Sgr A East is a mixed-morphology supernova remnant, which means in this case that it has a spherical radio shell surrounding a centrally concentrated X-ray continuum source. In addition, the 6.7 keV iron emission line is concentrated toward the center of the shell, consistent with Sgr A East being a supernova remnant. The implied high metallicity and the placement of the iron line emission is not accounted for by hypotheses invoking stellar disruption or an energy release from the stellar black hole.

One of the reasons that Sgr A East is said to be unusually energetic for a supernova remnant is that it has apparently compressed the dust and gas in the interstellar medium around it, particularly on the Eastern side of its periphery, where a ridge of dense molecular gas abuts the edge of the shell. This ridge is apparently the compressed portion of the 50 km s^{-1} molecular cloud (Serabyn *et al* 1992; Uchida *et al* 1998). Within this ridge, and aligned with the edge of the Sgr A East shell, is a string of compact H II regions, G-0.02-0.07, indicating that massive stars have recently formed within the compressed gas ridge. While one is tempted to conclude that the young stars responsible for these H II regions were formed as a result of the cloud compression by Sgr A East, there is a timescale problem: it takes substantially longer to form stars ($\sim 10^5$ years) than the expansion time of Sgr A East.

The question of whether the shell of Sgr A East has overrun the central black hole is an important one for understanding the environment of the black hole. The shell of Sgr A East coincides roughly with the inner edge of the CND on the side opposite to the center of Sgr A East (although that shell is not terribly well defined there), so it is entirely possible that the inner parsec of the Galaxy lies within Sgr A East, where the gas in most of the volume is extremely hot ($\sim 10^7 - 10^8$ K). The unknown quantity is the line-of-sight displacement between them. If the shell has passed over the black hole within the past several hundred years, as considered
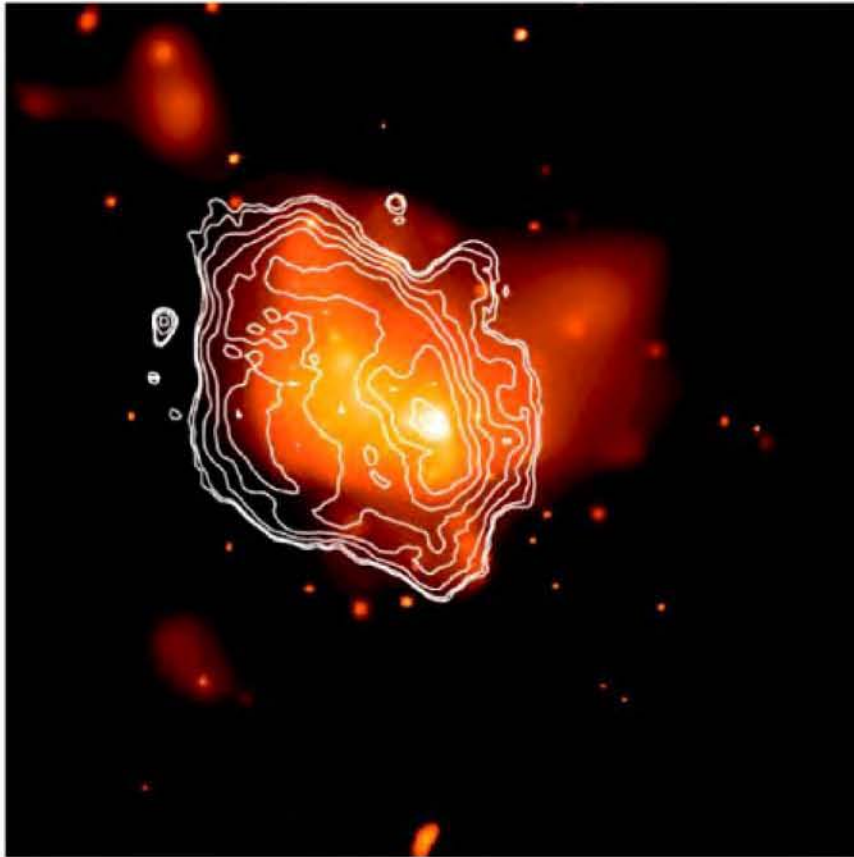
Figure 4.12. Smoothed X-ray image of Sgr A East (1.5–7.0 keV) with superimposed 20 cm radio contours, from Maeda *et al* (2002). See also color section.

by Maeda *et al* (2002), then because of the strongly enhanced density of the shell, this event may have led to an energetic accretion event. The dynamics of such an event, or of the gas presently surrounding the black hole if the shell has already swept through the center, are complicated by the strong ram pressure of the winds emanating from the hot stars in the central parsec. Indeed, most investigators assume that the black hole resides within the stellar wind bubble. However, the current Chandra X-ray picture, which shows continuum X-ray emission extended throughout this region, is consistent with the idea that much of the volume of the central parsec is filled with a hot gas.

4.9   The vicinity of Sgr A*

The immediate neighborhood of Sgr A* shows a modest amount of activity when examined closely, but displays nothing that would herald the presence of a supermassive black hole, except for a localized X-ray source. Sgr A* is not coincident with the plasma structures that constitute Sgr A West, although it lies at the end of a string of tiny, thermally emitting blobs which may have resulted from gas dynamics in the presence of a black hole (sources ϵ, ζ, and η in Yusef-Zadeh *et al* 1990; see also figure 4.10). Wardle and Yusef-Zadeh (1992) and Melia *et al* (1996) have suggested that these blobs were formed when the collective winds from the emission-line stars closest to Sgr A* (the stars constituting the IRS16 complex), which are all displaced to the East of it, pass by the black hole and are gravitationally concentrated downstream into the observed blobs. The fact that these are a series of blobs rather than a continuous stream can presumably be attributed in this scenario to a thermal instability in the compressed flow. Kinematical studies of the plasma blobs, including both proper motions and radial velocities, will ultimately enable this hypothesis to be tested. In any case, understanding these winds is extremely important because this is likely to be the source of matter accreted onto the black hole (chapter 10).

The region immediately around Sgr A* in projection also shows up as faint, extended, mid-infrared emission (Stolovy *et al* 1996; Morris *et al* 2001 and in preparation), although it cannot be associated directly with the black hole because such emission can be found over much of the region, probably as a result of thermal emission from warm dust in the Sgr A West complex. In fact, this lumpy, extended mid-infrared emission will make it difficult to measure the flux of Sgr A* at these crucial wavelengths above the synchrotron cut-off.

X-ray emission from Sgr A* has recently been measured with the CXO by Baganoff *et al* (2002). The steady-state source is relatively dim, with a 0.5–10 keV luminosity of 4×10^{33} erg s^{-1}, and appears to be extended by about 1 arcsec, or $\sim$0.04 pc. This persistent flux (constant over at least a few years) may be attributable to emission from the outermost parts of the accretion flow. During an observation with the CXO in 2000, Sgr A* underwent a few-hour flare, increasing its X-ray flux by a factor of about 50 relative to the quiescent value (Baganoff *et al* 2001). The short timescale ($\sim$20 min) of the substructure of this flare dictates that it must have come from a region less than about 20 Schwarzschild radii in size. The X-rays can therefore allow us to probe the environment of the black hole all the way in to near the event horizon. The theoretical interpretation of the flare emission is discussed by Markoff *et al* (2001; see also chapter 11). Ongoing simultaneous observations at a variety of wavelengths should really help constrain models for the events or instabilities which produced the X-rays.

4.10 Perspective

The study of the region about the Galactic Center did not begin in earnest until infrared and radio astronomy reached maturity in the 1960s and 1970s. Now, the tableau is well filled out, but we are left with many fascinating puzzles. The myriad ways in which the central black hole has been interacting with its environment are only just beginning to be appreciated. The future of research in this area will presumably reward us with clearer pictures of how matter migrates inward from the rest of the Galaxy and forms stars, of how the strong magnetic field can alter cloud dynamics and affect star formation and accretion into the central parsec, and how energy release from the black hole as it accretes gas or swallows stars can act back upon the surrounding medium. Fortunately, some events near the black hole happen quickly or often enough that we may be able to witness them in action (the orbital motions of the nearest stars, for example). However, much of the most interesting work facing researchers will require finding subtle clues to the energetic events that have taken place in the past.

References

Aitken D K, Smith C H, Moore T J T and Roche P F 1998 *Mon. Not. R. Astron. Soc.* **299** 743

Alexander T 1999 *Astrophys. J.* **527** 835

Baganoff F K *et al* 2001 *Nature* **413** 45

——2002 *Astrophys. J.* (astro-ph/0102151) submitted

Becklin E E, Gatley I and Werner M W 1982 *Astrophys. J.* **258** 135

Becklin E E and Neugebauer G 1968 *Astrophys. J.* **151** 145

Chandran B D G 2001 *Astrophys. J.* **562** 737

Chandran B D G, Cowley S C and Morris M 2000 *Astrophys. J.* **528** 723

Chandrasekhar S and Fermi E 1953 *Astrophys. J.* **118** 113

Davidson J A, Werner M W, Wu X, Lester D F, Harvey P M, Joy M and Morris M 1992 *Astrophys. J.* **387** 189

Dent W R F, Matthews H E, Wade R and Duncan W D 1993 *Astrophys. J.* **410** 650

Ebisuzaki T, Makino J, Tsuru T G, Funato Y, Portegies Zwart S, Hut P, McMillan S, Matsushita S, Matsumoto, H and Kawabe R 2001 *Astrophys. J. Lett.* **562** L19

Ekers R D, van Gorkom J H, Schwarz U J and Goss W M 1983 *Astron. Astrophys.* **122** 143

Figer D F, Becklin E E, McLean I S and Morris M 1994 *Infrared Astronomy with Arrays: the Next Generation* ed I S McLean (Dordrecht: Kluwer) p 545

Figer D F, Kim S S, Morris M, Serabyn E, Rich R M and McLean I S 1999b *Astrophys. J.* **525** 750

Figer D F, McLean I S and Morris M 1999a *Astrophys. J.* **514** 202

Gatley I, Jones J J, Hyland A R, Wade R and Geballe T R 1986 *Mon. Not. R. Astron. Soc.* **222** 299

Gaume R A, Claussen M J, De Pree C G, Goss W M and Mehringer D M 1995 *Astrophys. J.* **449** 663

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219

Genzel R, Pichon C, Eckart A, Gerhard O E and Ott T 2000 *Mon. Not. R. Astron. Soc.* **317** 348

Gerhard O 2001 *Astrophys. J. Lett.* **546** 39

Gezari S, Ghez A M, Becklin E E, Larkin J, McLean I S and Morris M 2002 *Astrophys. J.* at press

Gray A D, Nicholls J, Ekers R D and Cram L E 1995 *Astrophys. J.* **448** 164

Güsten R 1989 *The Center of the Galaxy (Proc. IAU Symp. 136)* ed M Morris (Dordrecht: Kluwer) p 89

Güsten R, Genzel R, Wright M C H, Jaffe D T, Stutzki J and Harris A I 1987 *Astrophys. J.* **318** 124

Herbst T M, Beckwith S W V and Forrest W J 1993 *Astrophys. J.* **105** 956

Hildebrand R H, Davidson J A, Dotson J, Figer D F, Novak G, Platt S R and Tao L 1993 *Astrophys. J.* **417** 565

Hildebrand R H, Gonatas D P, Platt S R, Wu X D, Davidson J A, Werner M W, Novak G and Morris M 1990 *Astrophys. J.* **362** 114

Ho L C and Filippenko A V 1996 *Astrophys. J.* **472** 600

Jackson J M, Geis N, Genzel R, Harris A I, Madden S, Poglitsch A, Stacey G J and Townes C H 1993 *Astrophys. J.* **402** 173

Kawai N, Fenimore E E, Middleditch J, Cruddace R G, Fritz G G, Snyder W A and Ulmer M P 1988 *Astrophys. J.* **330** 130

Khokhlov A and Melia F 1996 *Astrophys. J. Lett.* **457** L61

Kim S S, Figer D F, Lee H M and Morris M 2000 *Astrophys. J.* **545** 301

Kim S and Morris M 2001 *Astrophys. J.* **554** 1059

——2002 *Preprint*

Kim S S, Morris M and Lee H M 2000 *Astrophys. J.* **525** 228

Koyama K, Awaki H, Kunieda H, Takano S and Tawara Y 1989 *Nature* **339** 603

Koyama K, Maeda Y, Sonobe T, Takeshima T, Tanaka Y and Yamauchi S 1996 *Publications Astron. Soc. Japan* **48** 249

Krabbe A *et al* 1995 *Astrophys. J. Lett.* **447** L95

Kramer C, Staguhn J, Ungerechts H and Sievers A 1998 *The Central Regions of the Galaxy and Galaxies (Proc. IAU Symp. 184)* ed Y Sofue (Dordrecht: Kluwer) p 173

Lang C C, Goss W M and Morris M 2001 *Astrophys. J.* **121** 2681

——2002 *Astrophys. J.* at press

Lang C C, Morris M and Echevarria L 1999 *Astrophys. J.* **526** 727

Launhardt R, Zylka R and Mezger P G 2002 *Astron. Astrophys.* **384** 112

Lindqvist M, Habing H J and Winnberg A 1992 *Astron. Astrophys.* **259** 118

Lis D C, Menten K M, Serabyn E and Zylka R 1994 *Astrophys. J. Lett.* **423** L39

Liszt H S and Spiker R W 1995 *Astrophys. J. Suppl.* **98** 259

Lugten J B, Genzel R, Crawford M K and Townes C H 1986 *Astrophys. J.* **306** 691

Maeda Y *et al* 2002 *Astrophys. J.* **570** 671

Maoz D, Barth A J, Ho L C, Sternberg A and Filippenko A V 2001 *Astrophys. J.* **121** 3048

Markevitch M, Sunyaev R A and Pavlinsky M 1993 *Nature* **364** 40

Markoff S, Falcke H, Yuan F and Biermann P L 2001 *Astron. Astrophys.* **379** L13

Marshall J, Lasenby A N and Harris A I 1995 *Mon. Not. R. Astron. Soc.* **277** 594

McGary R S, Coil A L and Ho P T P 2001 *Astrophys. J.* **559** 326

Mehringer D M, Palmer P, Goss W M and Yusef-Zadeh F 1993 *Astrophys. J.* **412** 684

Melia F, Coker R F and Yusef-Zadeh F 1996 *Astrophys. J.* **460** 33

Mezger P G, Zylka R, Salter C J, Wink J E, Chini R, Kreysa E and Tuffs R 1989 *Astron.*

Astrophys. **209** 337

Morris M 1993 *Astrophys. J.* **408** 496

——1994 *The Nuclei of Normal Galaxies: Lessons from the Galactic Center (NATO ASI Series C vol 445)* ed R Genzel and A I Harris (Dordrecht: Kluwer) p 185

——1996 *Unsolved Problems in the Milky Way (Proc. IAU Symp. 169)* ed L Blitz and P J Teuben (Dordrecht: Kluwer) p 247

——1998 *The Central Regions of the Galaxy and Galaxies (Proc. IAU Symp. 184)* ed Y Sofue (Dordrecht: Kluwer) p 331

——2001 *Starburst Galaxies: Near and Far (Proc. in Physics)* ed L Tacconi and D Lutz (Berlin: Springer) p 53

Morris M, Davidson J A, Werner M, Dotson J, Figer D F, Hildebrand R, Novak G and Platt S 1992 *Astrophys. J. Lett.* **399** L63

Morris M, Ghez A M and Becklin E E 1999 *Adv. Spa. Res.* **23** 959

Morris M and Maillard J-P 2000 *Imaging the Universe in 3 Dimensions: Astrophysics with Advanced Multi-Wavelength Imaging Devices (ASP Conf. Proc. 195)* ed J Bland-Hawthorn and W van Breugel (New York: ASP) p 196

Morris M and Serabyn E 1996 *Annu. Rev. Astron. Astrophys.* **34** 645

Morris M, Tanner A M, Ghez A M, Becklin E E, Cotera A, Werner M W and Ressler M E 2001 *Bull. Am. Astron. Soc.* **198** 41.01

Najarro F, Krabbe A, Genzel R, Lutz D, Kudritzki R P and Hillier D J 1997 *Astron. Astrophys.* **325** 700

Novak G, Dotson J L, Dowell C D, Hildebrand R H, Renbarger T and Schleuning D A 2000 *Astrophys. J.* **529** 241

O'Connell R W, Gallagher J S and Hunter D A 1994 *Astrophys. J.* **433** 65

O'Connell R W, Gallagher J S, Hunter D A and Colley W N 1995 *Astrophys. J. Lett.* **446** L1

Paumard T, Maillard J-P and Morris M 2001 *Astron. Astrophys.* **366** 466

Pedlar A, Anantharamaiah K R, Ekers R D, Goss W M, van Gorkom J H, Schwarz U J and Zhao J-H 1989 *Astrophys. J.* **342** 769

Portegies-Zwart S F, Makino J, McMillan S L W and Hut P 2002 *Astrophys. J.* **565** 265

Roberts D A and Goss W M 1993 *Astrophys. J. Suppl.* **86** 133

Roberts D A, Yusef-Zadeh F and Goss W M 1996 *Astrophys. J.* **459** 627

Sanders R H 1998 *Mon. Not. R. Astron. Soc.* **294** 35

Serabyn E, Keene J, Lis D C and Phillips T G 1994 *Astrophys. J.* **424** L95

Serabyn E, Lacy J H and Achtermann J M 1992 *Astrophys. J.* **395** 166

Serabyn E, Lacy J H, Townes C H and Bharat R 1988 *Astrophys. J.* **326** 171

Serabyn E and Morris M 1994 *Astrophys. J.* **424** L91

——1996 *Nature* **382** 602

Shipman R F, Egan M P and Price S D 1997 *Galactic Center Newsletter* **5** 3

Sofue Y and Fujimoto M 1987 *Publications Astron. Soc. Japan* **39** 843

Staguhn J, Stutzki J, Uchida K I and Yusef-Zadeh F 1998 *Astron. Astrophys.* **336** 290

Stark A A, Gerhard O E, Binney J and Bally J 1991 *Mon. Not. R. Astron. Soc.* **248** 14

Stolovy S R, Hayward T L and Herter T 1996 *Astrophys. J. Lett.* **470** 45

Sunyaev R A, Markevitch M and Pavlinsky M 1993 *Astrophys. J.* **407** 606

Tacconi-Garman L E, Sternberg A and Eckart A 1996 *Astrophys. J.* **112** 918

Telesco C M, Davidson J A and Werner M W 1996 *Astrophys. J.* **456** 541

Tsuboi M, Handa T and Ukita N 1999 *Astrophys. J. Suppl.* **120** 1

Turner J L, Beck S C and Ho P T P 2000 *Astrophys. J. Lett.* **532** L109

Uchida K I, Morris M, Serabyn E, Fong D and Meseroll T 1998 *The Central Regions of the Galaxy and Galaxies (Proc. IAU Symp. 184)* ed Y Sofue (Dordrecht: Kluwer) p 317

Uchida K I, Morris M, Serabyn E and Güsten R 1996 *Astrophys. J.* **462** 768

Vollmer B and Duschl W J 2000 *New Astron.* **4** 581

——2001a *Astron. Astrophys.* **367** 72

——2001b *Astron. Astrophys.* **377** 1016

Wang Q D, Gotthelf E V and Lang C C 2002 *Nature* **415** 148

Wardle M and Yusef-Zadeh F 1992 *Nature* **357** 308

White G J 1996 *The Galactic Center (ASP Conf. Ser. 102)* ed R Gredel (New York: ASP) p 171

Wright M C H, Coil A L, McGary R S, Ho P T P and Harris A I 2001 *Astrophys. J.* **551** 254

Yamauchi S, Kawada M, Koyama K, Kunieda H and Tawara Y 1990 *Astrophys. J.* **365** 532

Yusef-Zadeh F and Morris M 1987a *Astrophys. J.* **94** 1178

——1987b *Astrophys. J.* **320** 545

Yusef-Zadeh F, Morris M and Chance D 1984 *Nature* **310** 557

Yusef-Zadeh F, Morris M and Ekers R D 1990 *Nature* **348** 45

Yusef-Zadeh F, Roberts D A, Goss W M, Frail D and Green A 1996 *Astrophys. J.* **466** L25

Yusef-Zadeh F, Stolovy S R, Burton M, Wardle M and Ashley M C B 2001 *Astrophys. J.* **560** 749

Yusef-Zadeh F, Stolovy S R, Burton M, Wardle M, Melia F, Lazio T J, Kassim N E and Roberts D A 1999 *The Central Parsecs of the Galaxy (ASP Conf. Ser. 186)* ed A Cotera, W J Duschl, F Melia and M J Rieke (New York: ASP) p 197

Zhao J-H, Desai K, Goss W M and Yusef-Zadeh F 1993 *Astrophys. J.* **418** 235

Zhao J-H and Goss W M 1998 *Astrophys. J.* **499** L163

This page intentionally left blank

PART 2

GENERAL RELATIVITY AND BLACK HOLES

This page intentionally left blank

Chapter 5

Particles and fields near black holes

Valeri Frolov
University of Alberta, Edmonton, Canada

Taking now the existence of black holes for granted, the motion of particles is studied in black hole spacetimes, first in the Schwarzschild (see chapter 1) and then in the Kerr background (see chapter 3). Subsequently, the propagation of fields in the same backgrounds is reviewed, taking a massless scalar field as a 'guinea pig'. Thereafter, more complicated spin-carrying fields are shortly discussed. Some physical effects, such as superradiance, are briefly mentioned. Finally, black hole electrodynamics is dealt with. A 3 + 1 decomposition of Maxwell's equations is carried out. The so-called membrane paradigm is introduced which treats the black hole as a black box with classical electrodynamic behavior. In this way, a black hole can serve as a kind of a dynamo. This mechanism may explain the activity of the nuclei of galaxies and quasars.

5.1 Introduction

A black hole is a region with a gravitational field so strong that no information-carrying signals can escape from this region to infinity. The gravitational field of a black hole which is formed in a non-spherical collapse is initially time dependent. But very soon (after the characteristic time $\sim 10^{-5}$ s $(M/M_\odot)$) it becomes stationary. After this an isolated black hole remains stationary or (if its angular momentum vanishes) static. Uniqueness theorems proved in the Einstein–Maxwell theory guarantee that the metric of a stationary black hole is uniquely specified by its mass M, angular momentum J, and electric charge Q. The charge usually does not play any important role in astrophysical applications. In the presence of surrounding matter and as a result of the accreted matter falling into the black hole, the parameters of stellar and supermassive black holes change slowly.

A black hole itself is an empty spacetime and by its very definition it is invisible[1]. Our conclusions concerning black hole existence are based on observations of matter falling into black holes. Matter falling from a distant region and having initially angular momentum cannot reach the black hole horizon until it loses the main part of its angular momentum. As a result, it forms an accretion disc.

Studying the dynamics of gas and plasma near stellar black holes and the motion of gas, plasma, and stars near massive black holes allows one to determine the mass of the object located in the central region of the system. If this mass is large enough, the only possibility to explain the dynamics of the system is to assume that the central object is a black hole.

It should be emphasized that this conclusion is based on the assumption that the general theory of relativity is valid. Most theoreticians believe that this is true. But up to now, most of the tests which confirm general relativity were performed in the weak-field limit. For this reason it is extremely important to obtain information from the very central regions (say five gravitational radii or smaller) of such objects. By studying details of the motion of matter and comparing the data with the predictions of general relativity, one can demonstrate that our understanding of the properties of spacetime in a strong gravitational field is correct. Under these conditions arguments based on the mass estimation of the central object practically leave no room for other options than a black hole.

To get templates which can be used for comparison with future astrophysical observations of very central regions, one must know solutions for the motion of plasma and particles in the vicinity of a black hole and for the electromagnetic (or gravitational) wave propagation from the central region to a distant observer. Both problems were studied in general relativity in great detail. In this chapter we collect the most important results concerning particle motion and field propagation in the black hole geometry. We focus our attention mainly on those effects which are characteristic for black holes.

In the first two sections we discuss particle motion in the Schwarzschild and the Kerr geometry. Field propagation in the vicinity of black holes is the subject of the third section. Black hole electrodynamics will be considered in the last section. Additional material connected with the subject of this chapter can be found in the books [1–6].

5.2 Particle motion near a non-rotating black hole

5.2.1 Equations of motion

5.2.1.1 Particle motion

Consider a body which has a size much smaller than the size of a black hole such that details connected with its internal structure are not important for the problem

---

[1] We do not consider here primordial black holes which can emit quantum Hawking radiation.

under consideration. Such an object is usually called a *particle*. Often we can consider as particles planet-like objects in their Keplerian motion near a black hole, small elements of accreting matter, stellar black holes and neutron stars falling into a massive black hole, and so on. We do not consider here the more complicated case when an object moving near a black hole has internal structure and its internal degrees of freedom can be excited by *tidal forces*.

Particle motion in the background black hole geometry is described by a solution of the *geodesic equation*

$$\frac{Du^\mu}{d\tau} := u^\nu u^\mu{}_{;\nu} = \frac{du^\mu}{d\tau} + \Gamma^\mu{}_{\nu\lambda} u^\nu u^\lambda = 0. \tag{5.1}$$

Here $u^\mu = dx^\mu/d\tau$ is the four-velocity of a particle, τ is the proper time and $\Gamma^\mu{}_{\nu\lambda}$ are the *Christoffel symbols*

$$\Gamma^\mu{}_{\alpha\beta} = g^{\mu\nu}\Gamma_{\nu\alpha\beta} \qquad \Gamma_{\nu\alpha\beta} = \tfrac{1}{2}(g_{\nu\alpha,\beta} + g_{\nu\beta,\alpha} - g_{\alpha\beta,\nu}). \tag{5.2}$$

5.2.1.2 Schwarzschild metric

The geometry of a static non-rotating black hole is spherically symmetric and described by the Schwarzschild metric

$$ds^2 = -\left(1 - \frac{r_S}{r}\right) dt^2 + \frac{dr^2}{1 - \frac{r_S}{r}} + r^2 d\Omega^2. \tag{5.3}$$

Here $r_S = 2M$ is the Schwarzschild gravitational radius, M is the black hole mass, and $d\Omega^2 = d\theta^2 + \sin^2\theta \, d\phi^2$ is the line element on the unit sphere. The gravitational radius r_S is the only essential dimensional parameter. It determines all characteristic time and length scales. The metric can be rewritten in the following form:

$$ds^2 = r_S^2\left[-\left(1 - \frac{1}{x}\right) d\tilde{t}^2 + \frac{dx^2}{1 - \frac{1}{x}} + x^2 d\Omega^2\right] \tag{5.4}$$

where $x := r/r_S$ is a dimensionless radial coordinate and $\tilde{t} := t/r_S$ a dimensionless time coordinate. Since the geodesic equations are scale invariant, it is sufficient to solve them for only one value of black hole mass (say for $r_S = 1$). The solutions for other masses can be obtained simply by rescaling.

Particle trajectories near a non-rotating black hole can be found by solving the geodesic equation in the Schwarzschild geometry. A more effective way is to use the *integrals of motion* connected with the spacetime *symmetries*.

5.2.2 Symmetries and integrals of motion

5.2.2.1 Killing vectors

A *Killing vector field* ξ^μ is a vector field which satisfies the *Killing* equation

$$\xi_{(\mu;\nu)} = 0. \tag{5.5}$$

A Killing vector is a generator of a symmetry transformation on a spacetime manifold. *Killing trajectories* are integral lines of the Killing vector field,

$$\frac{dx^\mu}{dt} = \xi^\mu. \tag{5.6}$$

If ξ^μ is regular, Killing trajectories form a foliation and one can introduce (at least locally) coordinates (t, y^i) where t is a parameter along the trajectory, and the y^i are constant on a given trajectory. It is easy to show that the metric $g_{\mu\nu}$ in these coordinates does not depend on t. This demonstrates explicitly the symmetry of the spacetime carrying a Killing vector.

5.2.2.2   Integrals of motion

For a geodesic motion the quantity $\xi_\mu u^\mu$ is constant along the particle world line and hence it is an *integral of motion*. This follows from

$$\frac{d}{d\tau}(\xi_\mu u^\mu) = \underbrace{\xi_{(\mu;\nu)}u^\mu u^\nu}_{=0,\ \text{see equation (5.5)}} + \underbrace{\xi_\mu u^\nu u^\mu{}_{;\nu}}_{=0:\ \text{see equation (5.1)}} = 0. \tag{5.7}$$

5.2.2.3   Symmetries of Schwarzschild spacetime

Being static and spherically symmetric, the Schwarzschild metric has four linearly independent Killing vector fields. One of them, in Schwarzschild coordinates,

$$\xi_{(t)} := \xi_{(t)}^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial t} \tag{5.8}$$

is connected with the time symmetry. The other three are generators of the three-parameter group of rotations preserving the geometry on a unit sphere:

$$\xi_1 := \xi_1^\mu \frac{\partial}{\partial x^\mu} = -\cos\phi \frac{\partial}{\partial\theta} + \cot\theta \sin\phi \frac{\partial}{\partial\phi}$$

$$\xi_2 := \xi_2^\mu \frac{\partial}{\partial x^\mu} = \sin\phi \frac{\partial}{\partial\theta} + \cot\theta \cos\phi \frac{\partial}{\partial\phi} \tag{5.9}$$

$$\xi_3 := \xi_{(\phi)} := \xi_{(\phi)}^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial\phi}.$$

The latter vector generates rotation around a z-axis passing through the north $\theta = 0$ and south $\theta = \pi$ poles of the black hole.

The Killing vector $\xi_{(t)}$ possesses the following properties:

- $\xi_{(t)}$ is orthogonal to the surface $t = $ constant. This property follows from the invariance of the metric with respect to time reflection $t \to -t$. (The metric is not only stationary but also *static*.)

- The *infinite redshift surface* where $\xi_{(t)}^2 = g_{tt} = 0$ coincides with a *Killing horizon*, which in its turn coincides with the *event horizon*, see later.

5.2.3 Equations of motion of a free test particle

5.2.3.1 *Orbits are planar*

Consider a particle moving in the Schwarzschild metric. Using spherical symmetry, one can always choose coordinates so that at the initial moment τ_0 one has $\theta_0 = \pi/2$ and $(d\theta/d\tau)|_0 = 0$. Since the only non-vanishing components of $\Gamma^\theta{}_{\mu\nu}$ are

$$\Gamma^\theta{}_{r\theta} = \frac{1}{r} \qquad \Gamma^\theta{}_{\phi\phi} = -\sin\theta\cos\theta \tag{5.10}$$

the θ-component of the geodesic equation of motion takes the form

$$\frac{d^2\theta}{d\tau^2} + \frac{1}{r}\frac{dr}{d\tau}\frac{d\theta}{d\tau} - \sin\theta\cos\theta\left(\frac{d\phi}{d\tau}\right)^2 = 0. \tag{5.11}$$

A solution of this equation for the given initial data is $\theta = \pi/2$. Thus, the trajectory of a particle is planar and we can assume it to lie in the equatorial plane $\theta = \pi/2$.

5.2.3.2 *Effective potential*

The Schwarzschild metric is invariant under time t and angular coordinate ϕ translations. The corresponding conserved quantities are ($\theta = \pi/2$):

$$\widetilde{E} = -\xi_{(t)\mu}u^\mu = \left(1 - \frac{r_{\rm S}}{r}\right)\frac{dt}{d\tau} \tag{5.12}$$

$$l = \xi_{(\phi)\mu}u^\mu = r^2\frac{d\phi}{d\tau}. \tag{5.13}$$

$\widetilde{E} = E/m$ is the specific energy of a particle (E being the energy, and m being the mass of the particle). The quantity $l = L/m$ is the specific angular momentum of a particle (L being its angular momentum). For the motion in the equatorial plane, the total angular momentum coincides with the azimuthal angular momentum. For a motion of the particle in the black hole exterior, both t and ϕ are monotonic functions of τ.

Using these relations and the normalization condition for the four-velocity u^μ

$$-1 = u_\mu u^\mu = g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} \tag{5.14}$$

one gets

$$\left(\frac{dr}{d\tau}\right)^2 = \widetilde{E}^2 - V(r) \tag{5.15}$$

where we introduced the *effective potential*

$$V(r) = \left(1 - \frac{r_{\rm S}}{r}\right)\left(\frac{l^2}{r^2} + 1\right). \tag{5.16}$$
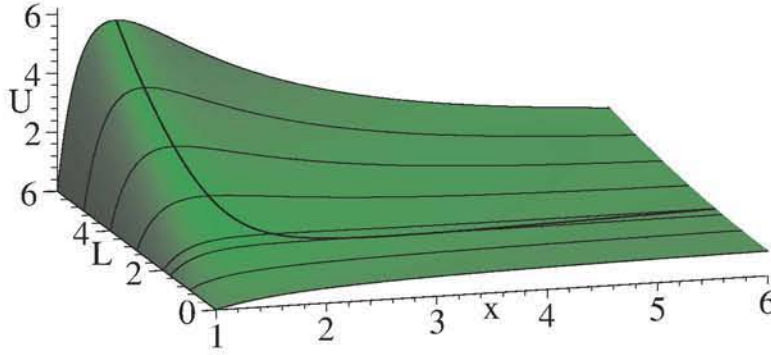
Figure 5.1. Effective potential $U(x)$ as a function of $x = r/r_S$ and for a specific angular momentum $\tilde{L}$ of a particle.

5.2.3.3 Properties of the effective potential

Different types of particle trajectory can be classified by studying the turning points of its radial motion where

$$V(r) = \widetilde{E}^2. \qquad (5.17)$$

The only scale parameter in the problem is the gravitational radius of the black hole. Using again the dimensionless coordinate x and $\tilde{L} = l/r_S$ we can rewrite $V(x)$ according to

$$V(r) = U(x) = \left(1 - \frac{1}{x}\right)\left(\frac{\tilde{L}^2}{x^2} + 1\right). \qquad (5.18)$$

The effective potential $U(x)$ is shown in figure 5.1. For fixed $|\tilde{L}| \geq \sqrt{3}$, U as a function of x has a maximum at x_+ and a minumum at x_-, where

$$x_\pm = \tilde{L}\left(\tilde{L} \pm \sqrt{\tilde{L}^2 - 3}\right). \qquad (5.19)$$

For $\tilde{L} = \sqrt{3}$, $x_- = x_+ = 3$. A heavy full line in figure 5.1 shows the position of the extrema.

5.2.4 Types of trajectory

5.2.4.1 Bound and unbound trajectories

The specific energy of a moving particle remains constant; in figure 5.2 this motion is shown by a horizontal line. The intersection of the horizontal line
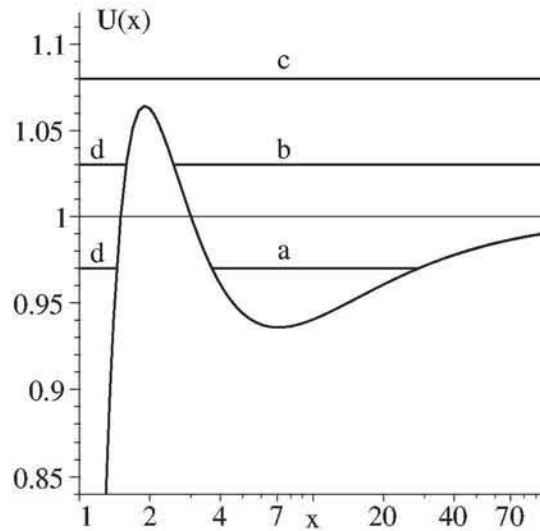
Figure 5.2. An example of the effective black hole potential for $\tilde{L} > \sqrt{3}$.

with the effective potential determines the turning points. The horizontal a with $\tilde{E}_1 < 1$ corresponds to the motion in a bound region in space between r_1 and r_2; this is an analogue of elliptic motion in Newtonian theory. The corresponding trajectory is not a conic section, and, in general, is not closed. If the orbit as a whole lies far from the black hole, it is an ellipse which slowly rotates in the plane of motion.

The segment b with $\tilde{E}_2 > 1$ corresponds to a particle coming from infinity and then moving back to infinity (an analogue of hyperbolic motion). Finally, the segment c with $\tilde{E}_3$ does not intersect the potential curve but passes above its maximum $\tilde{E}_{max}$. It corresponds to a particle falling into the black hole (*gravitational capture*). This type of motion is impossible in Newtonian theory and is typical for the black hole. Gravitational capture becomes possible because the effective potential has a maximum. No such maximum appears in the effective potential of Newtonian theory.

In addition, another type of motion is possible in the neighborhood of a black hole. This line d (with energy $\tilde{E}_4$) may lie below or above unity, stretching from r_S to the intersection with the curve $V(r)$. This segment represents the motion of a particle which, for example, first recedes from the black hole and reaches r_{max} (at the point of intersection of $\tilde{E}_4$ and $V(r)$), and then again falls toward the black hole and is absorbed by it. Examples of different types of trajectory are shown in figure 5.3.
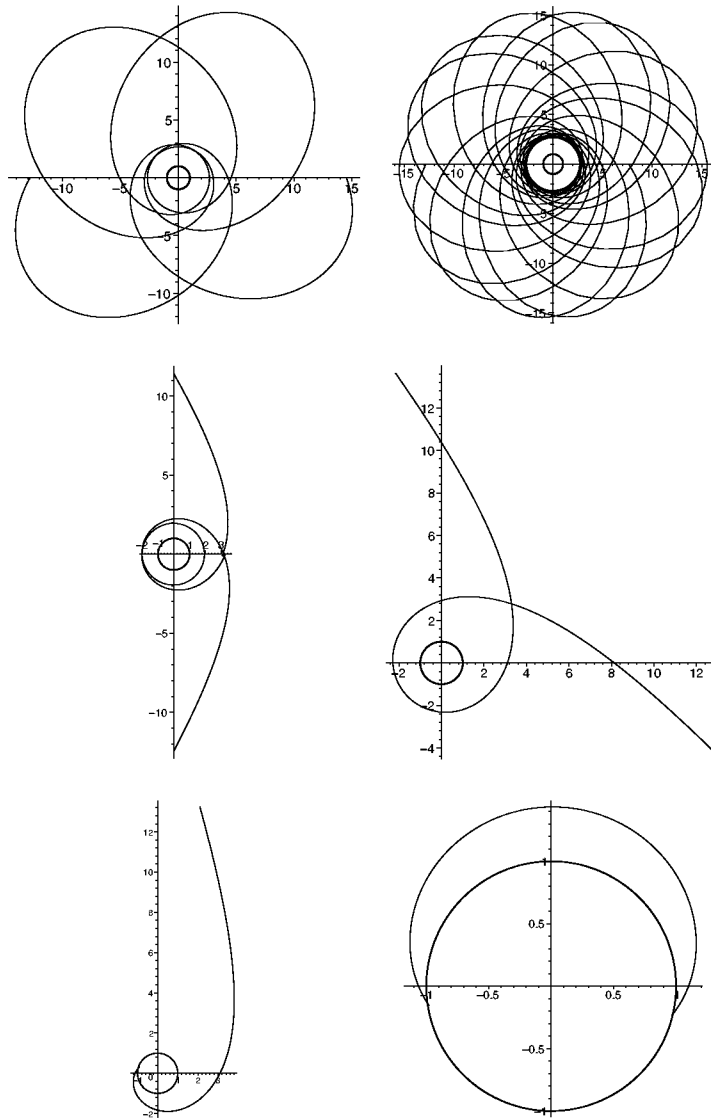
Figure 5.3. Different types of particle trajectory.

A body can escape to infinity if $\widetilde{E} \geq 1$. From equation (see [6, p 40])

$$\widetilde{E}^2 = (1 - r_S/r)(1 - v^2/c^2)^{-1} = 1 \qquad (5.20)$$
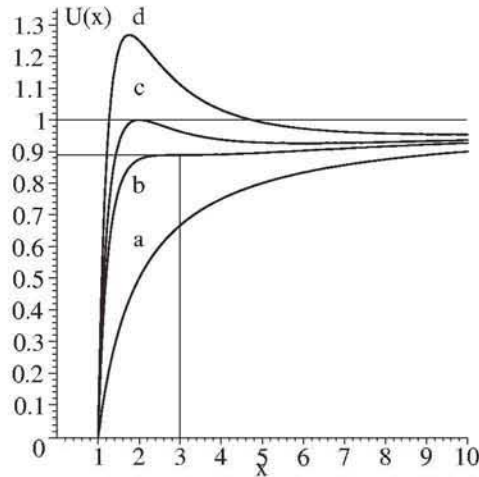
Figure 5.4. Effective black hole potential for different values of $\tilde{L}$: (a) $\tilde{L} = 0$, (b) $\tilde{L} = \sqrt{3}$, (c) $\tilde{L} = 2$, and (d) $\tilde{L} = \sqrt{6}$.

we find the escape velocity as

$$v_{esc} = c\sqrt{r_S/r} = \sqrt{2GM/r} \tag{5.21}$$

which coincides with the Newtonian expression. Note that in Newtonian theory in the gravitational field of a pointlike mass, the escape velocity guarantees the escape to infinity regardless of the direction of motion. The case of the black hole is different. Even if a particle has the escape velocity, it can be trapped by the black hole, the latter occurring if the particle moves towards the black hole. We have already mentioned this effect, calling it *gravitational capture*.

5.2.4.2   Circular motion

For circular motion around a black hole $dr/d\tau \equiv 0$. This motion is represented in figure 5.2 by a point at the extremum of the effective potential curve. A point at the minimum corresponds to a stable motion, and a point at the maximum to an unstable motion. The latter motion has no analogue in Newtonian theory and is specific to black holes. If the motion of a particle is represented by a horizontal line $\widetilde{E} = $ constant very close to $\widetilde{E}_{max}$, then the particle makes many turns around the black hole at a radius close to r corresponding to $\widetilde{E}_{max}$ before the orbit moves far away from this value of r. The shape and the position of the potential $V(r)$ are different for different $\tilde{L}$: the corresponding curves for some values of $\tilde{L}$ are shown in figure 5.4.

The maximum and minimum appear on $V(r)$ curves when $\tilde{L} > \sqrt{3}$. If $\tilde{L} < \sqrt{3}$ the $V(r)$ curve is monotone. Hence, the motion on circular orbits is possible only if $\tilde{L} > \sqrt{3}$. The minima of the curves then lie at $r > 3r_{\mathrm{S}}$. Stable circular orbits thus exist only for $r > 3r_{\mathrm{S}}$. At smaller distances, there are only unstable circular orbits corresponding to the maximum of $\tilde{E}_{\max}$ curves. If $\tilde{L} \to \infty$, the position of the maximum of the $\tilde{E}_{\max}$ curve decreases to $r = 1.5r_{\mathrm{S}}$. Even unstable inertial circular motion becomes impossible at r less than $1.5r_{\mathrm{S}}$.

The critical circular orbit that separates stable motions from unstable ones corresponds to $r = 3r_{\mathrm{S}}$. Particles move along it at a velocity $v = c/2$, the energy of a particle being $\tilde{E} = \sqrt{8/9} \approx 0.943$. This is the motion with the maximum possible binding energy $E \approx 0.057\ mc^2$.

Let us emphasize the importance of this result for black hole astrophysics. Suppose a non-rotating black hole is surrounded by a thin accretion disk. Let us follow the time evolution of a matter element of the disc. It is moving along a practically circular orbit slowly losing its energy and angular momentum until it reaches the position of the last stable circular orbit. After this, it falls almost freely into the black hole. This means that the maximum efficiency of the energy release by matter falling into a non-rotating black hole is 5.7%.

The velocity on (unstable) orbits, with $r < 3r_{\mathrm{S}}$, increases as r decreases from $c/2$ to c on the last circular orbit with $r = 1.5r_{\mathrm{S}}$. When $r = 2r_{\mathrm{S}}$, the particle's energy is $\tilde{E} = 1$, that is, the circular velocity is equal to the *escape velocity*. If r is still smaller, the escape velocity is smaller than the circular velocity. There is no paradox in it, since the circular motion here is unstable and even the tiniest perturbation (supplying momentum away from the black hole) transfers the particle to an orbit moving it to infinity, that is, an orbit corresponding to hyperbolic motion.

5.2.5 Equations of motion in 'tilted' spherical coordinates

To simplify the equations of motion of a test particle we used a special choice of coordinates, namely we oriented the z-axis (that is the direction $\theta = 0, \pi$) to be orthogonal to the plane of the orbit. Let us now check how the equations of motion are modified if the z-axis is tilted and not orthogonal to the orbit plane. This exercise is instructive for the discussion of particle motion in the Kerr geometry where there exists a preferred direction of the z-axis determined by the direction of the angular momentum of the rotating black hole.

The expression for $\tilde{E}$ remains the same, while the specific azimuthal angular momentum, which we denote now l_z, is

$$l_z = r^2 \sin^2 \theta \frac{d\phi}{d\tau}.$$ (5.22)

One also needs the expression for the conserved *total angular momentum*, l,

$$l^2 = r^4 \left[\left(\frac{d\theta}{d\tau} \right)^2 + \sin^2 \theta \left(\frac{d\phi}{d\tau} \right)^2 \right].$$ (5.23)

Using these relations and the normalization condition $u_\mu u^\mu = -1$ one can obtain the following set of equations:

$$\frac{dr}{d\tau} = \pm \left[\widetilde{E}^2 - \left(1 - \frac{r_S}{r}\right)\left(\frac{l^2}{r^2} + 1\right)\right]^{1/2} \tag{5.24}$$

$$r^2\frac{d\theta}{d\tau} = \pm \left(l^2 - \frac{l_z^2}{\sin^2\theta}\right)^{1/2} \tag{5.25}$$

$$r^2\frac{d\phi}{d\tau} = \frac{l_z}{\sin^2\theta} \tag{5.26}$$

$$\frac{dt}{d\tau} = \frac{\widetilde{E}}{1 - \frac{r_S}{r}}. \tag{5.27}$$

The equation for $\theta(\tau)$ shows that the angle θ changes between θ_0 and $\pi - \theta_0$, where $\sin\theta_0 = l_z/l$. This means that the angle between the normal to the trajectory plane and z-axis is $\pi/2 - \theta_0$.

5.2.6 Motion of ultrarelativistic particles

5.2.6.1 Equations of motion

When the energy E is much larger than m, a particle is called ultrarelativistic. In this limit $\widetilde{E} \to \infty$ and $\widetilde{L} \to \infty$ while the ratio $\widetilde{L}/\widetilde{E}$ remains finite and is equal to $\tilde{b} := b/r_S$, where b is the impact parameter of the particle at infinity. The equations of motion of the ultrarelativistic particle (or a light ray) take the form ($\tilde{t} = t/r_S$):

$$\left(\frac{dx}{d\tilde{t}}\right)^2 = \left(1 - \frac{1}{x}\right)^2\left[1 - \frac{\tilde{b}^2}{x^2}\left(1 - \frac{1}{x}\right)\right] \tag{5.28}$$

$$\frac{d\phi}{d\tilde{t}} = \left(1 - \frac{1}{x}\right)\frac{\tilde{b}}{x^2}. \tag{5.29}$$

The sign of b depends on the sense of motion; we assume that b is positive. The radial turning point on the trajectory is defined by the equation

$$x^3 - \tilde{b}^2(x - 1) = 0. \tag{5.30}$$

The impact parameter b as a function of the position of a radial turning point is shown in figure 5.5.

5.2.6.2 Types of trajectory

In figure 5.5, the motion of an ultrarelativistic particle with a given b is represented by a horizontal line $b = $ constant. A particle approaches the black hole,
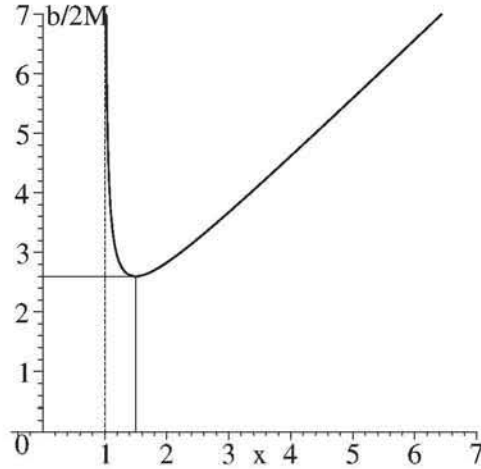
Figure 5.5. Impact parameter b as a function of the position of extrema in $x = r/r_S$ on the trajectory of an ultrarelativistic particle.

passes by it at the minimal distance corresponding to the point of intersection of b = constant with the right-hand branch of the $b(r)$ curve, and again recedes to infinity. If the intersection occurs close to the minimum $b_{min} = 3\sqrt{3} \times r_S/2$, the particle may experience a number of turns before it flies away to infinity. The exact minimum of the curve $b(r)$ corresponds to the (unstable) motion on a circle of radius $r = 1.5r_S$ at the velocity $v = c$. Note that the left-hand branch of $b(r)$ in figure 5.5 corresponds to the maximum distance between the ultrarelativistic particle and the black hole; the particle first recedes to $r < 1.5r_S$ but then again falls into the black hole. Obviously, for such a motion the parameter b does not have the literal meaning of the impact parameter at infinity since the particle never recedes to infinity. For a given coordinate r, this parameter can be found as a function of the angle ψ between the trajectory of the particle and the direction to the center of the black hole:

$$b = \frac{r|\tan\psi|}{\sqrt{(1 - r_S/r)(1 + \tan^2\psi)}}. \tag{5.31}$$

If an ultrarelativistic particle approaches the black hole on the way from infinity and the parameter b is less than the critical value $b_{min} = 3\sqrt{3}r_S/2$, the particle falls into the black hole.

5.2.7   Gravitational capture

Let us consider now the motion of a test particle in which its trajectory terminates in the black hole. Two types of such a motion are possible. First, the trajectory of the particle starts at infinity and ends in the black hole. Second, the trajectory starts and ends in the black hole. Of course, a particle cannot be ejected from the black hole. Hence, the motion on the second-type trajectory becomes possible either if the particle was placed on this trajectory via a non-geodesic curve or if the particle was created close to the black hole.

The gravitational capture of a particle coming from infinity is of special interest. Let us have a better look at this case. It is clear from the analysis of motion given in the preceding section that a particle coming from infinity can be captured if its specific energy $\widetilde{E}$ is greater, for a given $\widetilde{L}$, than the maximum $(\widetilde{E}_{\max})$ of the curve $V(r)$. Let us consider the gravitational capture in two limiting cases, one for a particle whose velocity at infinity is much lower than the speed of light ($v_\infty/c \ll 1$) and another for a particle which is ultrarelativistic at infinity.

In the former case, $\widetilde{E} \approx 1$. The curve $V(r)$, which has $\widetilde{E}_{\max} = 1$, corresponds to $\widetilde{L}_{\mathrm{cr}} = 2$ (line c in figure 5.2). The maximum of this curve lies at $r = 2r_{\mathrm{S}}$. This radius is minimal for the periastra of the orbits of the particles with $v_\infty = 0$ which approach the black hole and again recede to infinity. If $\widetilde{L} \le 2$, gravitational capture takes place. The angular momentum of a particle moving with the velocity v_∞ at infinity is $L = m v_\infty b$, where b is the impact parameter. The condition $\widetilde{L} \equiv L/mcr_{\mathrm{S}} = 2$ defines the critical value $b_{\mathrm{cr,nonrel}} = 2r_{\mathrm{S}}(c/v_\infty)$ of the impact parameter for which the capture takes place. The capture cross section for a non-relativistic particle is

$$\sigma_{\mathrm{nonrel}} = \pi b_{\mathrm{cr}}^2 = 4\pi (c/v_\infty)^2 r_{\mathrm{S}}^2. \tag{5.32}$$

For an ultrarelativistic particle, $b_{\mathrm{cr}} = 3\sqrt{3}r_{\mathrm{S}}/2$, and the capture cross section is

$$\sigma_{\mathrm{rel}} = \frac{27}{4}\pi r_{\mathrm{S}}^2. \tag{5.33}$$

Owing to a possible gravitational capture, not every particle whose velocity exceeds the escape limit flies away to infinity. In addition, it is necessary that the angle ψ between the direction to the black hole center and the trajectory be greater than a certain critical value ψ_{cr}. For the velocity equal to the escape threshold this critical angle is given by the expression

$$\tan \psi_{\mathrm{cr,esc}} = \pm \frac{2\sqrt{(1 - r_{\mathrm{S}}/r)r_{\mathrm{S}}/r}}{\sqrt{1 - 4r_{\mathrm{S}}/r(1 - r_{\mathrm{S}}/r)}}. \tag{5.34}$$

The plus sign is chosen for $r > 2r_{\mathrm{S}}$ ($\psi_{\mathrm{cr}} < 90°$), and the minus sign for $r < 2r_{\mathrm{S}}$ ($\psi_{\mathrm{cr}} > 90°$).

For an ultrarelativistic particle, the critical angle is given by the formula

$$\tan \psi_{\mathrm{cr,rel}} = \pm \frac{\sqrt{1 - r_{\mathrm{S}}/r}}{\sqrt{r_{\mathrm{S}}/r - 1 + \frac{4}{27}(r/r_{\mathrm{S}})^2}}. \tag{5.35}$$

The plus sign is taken for $r > 1.5r_S$ and the minus for $r < 1.5r_S$.

5.3   Particle motion near a rotating black hole

5.3.1   Gravitational field of a rotating black hole

5.3.1.1   Kerr metric

If a black hole is rotating, the direction of its axis of rotation singles out a preferred direction in space. As a result, spherical symmetry characterizing the spacetime of a non-rotating black hole is broken. The geometry of a rotating black hole is axisymmetric. The *Kerr metric* describing this geometry written in the coordinates proposed by Boyer and Lindquist [7] is

$$ds^2 = -\left(1 - \frac{2Mr}{\Sigma}\right) dt^2 - \frac{4Mra\sin^2\theta}{\Sigma} dt\,d\phi + \frac{\Sigma}{\Delta} dr^2 + \Sigma\,d\theta^2 + \frac{A\sin^2\theta}{\Sigma} d\phi^2 \tag{5.36}$$

where

$$\Sigma \equiv r^2 + a^2\cos^2\theta \qquad \Delta \equiv r^2 - 2Mr + a^2$$
$$A = (r^2 + a^2)^2 - a^2\Delta\sin^2\theta. \tag{5.37}$$

Two constants which enter the Kerr metric are the black hole mass M and the rotation parameter a connected with angular momentum J of the black hole, $a = J/M$. As for the Schwarzschild metric, one can always rewrite the Kerr metric in the dimensionless form, by extracting the scale parameter of length dimensions from the metric. Since the radius r_+ of the event horizon depends on a, it is convenient to use $2M$ or M as the scaling parameter. The latter option is used more often. Note that in this case the normalization is different from the one used in the Schwarzschild case by an extra factor of two. The only dimensionless parameter which enters the dimensionless form of the Kerr metric is a/M, which can take values in the range $(-1, 1)$. For $|a/M| > 1$ there is no event horizon and the metric describes a spacetime with a naked singularity.

5.3.1.2   Killing vectors

Being stationary (independent of time t) and axisymmetric (independent of an angular coordinate ϕ) the Kerr metric has two Killing vectors:

$$\xi_{(t)} := \xi_{(t)}^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial t} \qquad \xi_{(\phi)} := \xi_{(\phi)}^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial \phi}. \tag{5.38}$$

The Kerr geometry and its Killing vectors possess the following properties:

- Since the component $g_{t\phi}$ of the metric does not vanish, the Killing vector field $\xi_{(t)}$ is tilted with respect to the section $t = $ constant. The tilting angle depends on r and θ.

- The *infinite redshift surface* where $\xi_{(t)}^2 \equiv g_{tt} = 0$ does not coincide with a *Killing horizon*. This surface where

$$r = M + \sqrt{M^2 - a^2 \cos^2 \theta} \qquad (5.39)$$

is an external boundary of the *ergosphere*.
- The event horizon lies at $\Delta = 0$, that is at $r = r_+$, where

$$r_+ = M + \sqrt{M^2 - a^2}. \qquad (5.40)$$

- The event horizon again coincides with the Killing horizon determined by the equation

$$\eta^2 = 0 \qquad (5.41)$$

where $\eta^\mu = \xi_{(t)}^\mu + \Omega^H \xi_{(\phi)}^\mu$, and

$$\Omega^H = \frac{a}{r_+^2 + a^2} \qquad (5.42)$$

is the *angular velocity* of the black hole.
- The infinite redshift surface lies everywhere outside the event horizon except at the two poles $\theta = 0$ and $\theta = \pi$. Inside the ergosphere, that is, between the infinite redshift surface and the horizon, the Killing vector field $\xi_{(t)}$ is spacelike, $\xi_{(t)}^2 > 0$.

5.3.1.3 Killing tensor

The dragging effect connected with the rotation of the black hole affects orbits of test particles. As a result, only orbits lying in the equatorial plane are planar. Two integrals of motion connected with two Killing vectors together with the proper time normalization condition are sufficient to reduce the equation of motion to a complete set of first integrals. To proceed with non-equatorial orbits an additional integral of motion is required. Fortunately such an integral exists for the Kerr geometry. It is connected with a *Killing tensor*.

A Killing tensor is a symmetric tensor field $\xi_{\mu\nu}$ obeying the equation

$$\xi_{(\mu\nu;\lambda)} = 0. \qquad (5.43)$$

In the same manner as for the Killing vector, one can show that for a geodesic motion the quantity

$$\mathcal{K} = \xi_{\mu\nu} u^\mu u^\nu \qquad (5.44)$$

remains constant along the worldline. Indeed

$$\frac{d}{d\tau}(\xi_{\mu\lambda} u^\mu u^\lambda) = \underbrace{\xi_{(\mu\lambda;\nu)} u^\mu u^\nu}_{=0:\text{see }(5.43)} u^\lambda + \xi_{\mu\lambda} u^\lambda \underbrace{u^\nu u^\mu_{;\nu}}_{=0:\text{ see }(5.1)} + \xi_{\mu\lambda} u^\mu \underbrace{u^\nu u^\lambda_{;\nu}}_{=0:\text{ see }(5.1)} = 0.$$

$$(5.45)$$

It is easy to check that a tensor product $\xi_\mu^{(1)}\xi_\nu^{(2)}$ of two Killing vectors $\xi_\mu^{(1)}$ and $\xi_\nu^{(2)}$ is a Killing tensor. In this case the corresponding conserved quantity is a product of two integrals of motion of the Killing vectors. A non-trivial conservation law is connected only with a Killing tensor which is linearly independent of tensor products of the Killing vectors. For the Kerr metric, such an independent Killing tensor has the following non-vanishing components in the Boyer–Lindquist coordinates [8–12]

$$\xi_{00} = a^2 \left[1 - \frac{2Mr\cos^2\theta}{\Sigma} \right] \qquad \xi_{11} = -\frac{a^2\cos^2\theta\,\Sigma}{\Delta}$$

$$\xi_{22} = r^2\Sigma \qquad \xi_{03} = -\frac{a\sin^2\theta}{\Sigma}[\Delta a^2\cos^2\theta + r^2(r^2+a^2)] \qquad (5.46)$$

$$\xi_{33} = \frac{\sin^2\theta}{\Sigma}[r^2(r^2+a^2)^2 + \tfrac{1}{4}\Delta a^4\sin^2 2\theta].$$

5.3.2 Equations of motion of a free test particle

5.3.2.1 Integrals of motion

Conserved quantities connected with Killing vectors $\xi_{(t)}$ and $\xi_{(\phi)}$ are:

$$\widetilde{E} = -\xi_{(t)\mu}u^\mu = \left(1 - \frac{2Mr}{\Sigma}\right)\frac{dt}{d\tau} + \frac{2Mra\sin^2\theta}{\Sigma}\frac{d\phi}{d\tau} \qquad (5.47)$$

$$l_z = \xi_{(\phi)\mu}u^\mu = -\frac{2Mra\sin^2\theta}{\Sigma}\frac{dt}{d\tau} + \frac{A\sin^2\theta}{\Sigma}\frac{d\phi}{d\tau}. \qquad (5.48)$$

As before, $\widetilde{E} = E/m$ is the specific energy and $l_z = L_z/m$ is the specific angular momentum of a particle. A conserved quantity connected with the Killing tensor is

$$\mathcal{K} = \left(\widetilde{E}a\sin\theta - \frac{l_z}{\sin\theta}\right)^2 + \Sigma^2\left(\frac{d\theta}{d\tau}\right)^2 + a^2\cos^2\theta. \qquad (5.49)$$

Quite often, one uses, instead of $\mathcal{K}$, another integral of motion, $\mathcal{Q}$, that is related to it by

$$\mathcal{Q} \equiv \mathcal{K} - (\widetilde{E}a - l_z)^2 = l_z^2\cot^2\theta - \widetilde{E}^2a^2\cos^2\theta + \Sigma^2\left(\frac{d\theta}{d\tau}\right)^2 + a^2\cos^2\theta. \qquad (5.50)$$

To summarize, the equations of motion of a particle in the Kerr–Newman spacetime allow four integrals of motion, E, L_z, $\mathcal{K}$ (or $\mathcal{Q}$), and a trivial one, $u^\mu u_\mu = -1$.

5.3.2.2 First integrals of the equations of motion

One can express the four components u^μ of the velocity as explicit functions of these integrals of motion and coordinates r and θ. As a result one gets the system

$$\Sigma \frac{dr}{d\tau} = \pm \mathcal{R}^{1/2} \tag{5.51}$$

$$\Sigma \frac{d\theta}{d\tau} = \pm \Theta^{1/2} \tag{5.52}$$

$$\Sigma \frac{d\phi}{d\tau} = \frac{l_z}{\sin^2 \theta} - a\widetilde{E} + \frac{a}{\Delta}[\widetilde{E}(r^2 + a^2) - l_z a] \tag{5.53}$$

$$\Sigma \frac{dt}{d\tau} = a(l_z - a\widetilde{E} \sin^2 \theta) + \frac{r^2 + a^2}{\Delta}[\widetilde{E}(r^2 + a^2) - l_z a] \tag{5.54}$$

where

$$\mathcal{R} = [\widetilde{E}(r^2 + a^2) - l_z a]^2 - \Delta[r^2 + (l_z - a\widetilde{E})^2 + \mathcal{Q}] \tag{5.55}$$

$$\Theta = \mathcal{Q} - \cos^2 \theta \left[a^2(1 - \widetilde{E}^2) + \frac{l_z^2}{\sin^2 \theta} \right]. \tag{5.56}$$

The signs $\pm$ which enter these relations are independent from one another.

In the limit $a \to 0$, that is for a non-rotating black hole, these equations coincide with the corresponding equations of motion in the tilted spherical coordinates. In this limit $\mathcal{Q} = l^2 - l_z^2$.

5.3.2.3 Bound and unbound motion

The geodesic world line of a particle in the Kerr metric is completely determined by the first integrals of motion $\widetilde{E}$, l_z, and $\mathcal{Q}$. Consider $\mathcal{R}$ which enters the radial equation of motion as a function of r for fixed values of the other parameters:

$$\mathcal{R} = (\widetilde{E}^2 - 1)r^4 + 2Mr^3 + [(\widetilde{E}^2 - 1)a^2 - l_z^2 - \mathcal{Q}]r^2$$
$$+ 2M[\mathcal{Q} + (\widetilde{E}a - l_z)^2]r - a^2\mathcal{Q}. \tag{5.57}$$

The leading term for large r on the right-hand side is positive if $\widetilde{E}^2 > 1$. Only in this case can the motion be infinite. For $E^2 < 1$ the motion is always finite, i.e. the particle cannot reach infinity.

5.3.2.4 Effective potential

For a rotating black hole the variety of trajectories becomes wider and their classification is much more involved [13, 14]. We discuss only some important classes of trajectories.

For studying the qualitative characteristics of the motion of test particles in the Kerr metric it is convenient to use the *effective potential*. Let us rewrite $\mathcal{R}$ as

$$\mathcal{R} = \alpha \widetilde{E}^2 - 2\beta \widetilde{E} + \gamma \tag{5.58}$$

where

$$\alpha = r^4 + a^2(r^2 + 2Mr) \qquad \beta = 2aMl_z r \qquad (5.59)$$

$$\gamma = l_z^2 a^2 - (r^2 + l_z^2 + Q)\Delta. \qquad (5.60)$$

The radial turning points $\mathcal{R} = 0$, see (5.51), are determined by the condition $\widetilde{E} = V_\pm(r)$, where

$$V_\pm := \frac{\beta \pm \sqrt{\beta^2 - \alpha\gamma}}{\alpha}. \qquad (5.61)$$

The quantities $V_\pm$ are known as the effective potentials. They are functions of r, the integrals of motion l_z and Q, and the parameters M and a. Actually, these quantities enter V only in the form of the dimensionless combinations r/M, l_z/M, Q/M^2, and a/M.

The motion of a particle with specific energy $\widetilde{E}$ is possible only in the regions where either $\widetilde{E} \geq V_+$ or $\widetilde{E} \leq V_-$. The function for $\mathcal{R}$ remains invariant under transformations $\widetilde{E} \to -\widetilde{E}$, $l_z \to -l_z$ relating the regions mentioned earlier. In the Schwarzschild geometry, the second region $\widetilde{E} \leq V_-$ is excluded, since, in the exterior of the black hole, $\widetilde{E} \geq 0$ and $V_- < 0$. The limiting values of the effective potentials $V_\pm$ at infinity and at the horizon respectively are:

$$V_\pm(r = \infty) = \pm 1 \qquad V_\pm(r_+) = al_z/2Mr_+ = \Omega^H l_z \qquad (5.62)$$

where Ω^H is the angular velocity of the black hole. The effective potentials for non-rotating and rapidly rotating black holes are shown in figure 5.6.

5.3.2.5   Motion in the θ-direction

Let us consider the properties of the function Θ which determines the motion of a particle in the θ-direction. Since $\Theta \geq 0$ the finite motion with $\widetilde{E}^2 < 1$ is possible only if $Q \geq 0$. The orbit is characterized by the value $Q = 0$ if and only if it is restricted to the equatorial plane. Non-equatorial *finite* orbits with $\theta = $ constant do not exist in the Kerr metric.

For $Q = 0$, Θ is positive only if $\widetilde{E}^2 > 1$. The turning points $\pm\theta_0$ in the θ-direction are defined by the equation

$$\sin^2 \theta_0 = \frac{l_z^2}{a^2(\widetilde{E}^2 - 1)}. \qquad (5.63)$$

This equation implies that $|l_z| \leq a\sqrt{\widetilde{E}^2 - 1}$. Since in this case all the coefficients which enter $\mathcal{R}$ are non-negative, there are no turning points in r. The corresponding motion is infinite. It starts either at infinity and ends at the black hole horizon, or it starts near the black hole horizon and ends at infinity.

For $Q \geq 0$, there exist both finite as well as infinite trajectories. They intersect the equatorial plane or (for $Q = 0$ and $\widetilde{E}^2 < 1$) are entirely situated in it. The particles with $Q < 0$ never cross the equatorial plane and move between two surfaces $\theta = \theta_+$ and $\theta = \theta_-$.
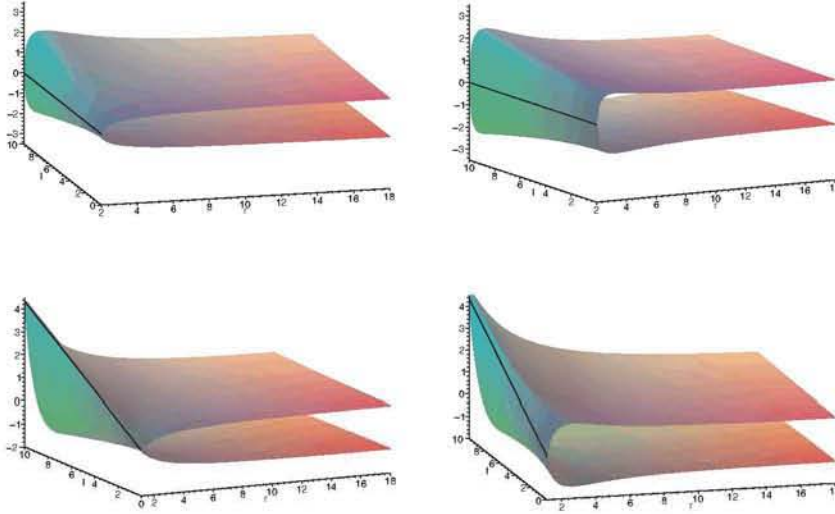
Figure 5.6. Effective potentials $V_\pm$ for the Kerr metric. The upper plots are for $a = 0$ ($Q = 0$ left and $Q = 40$ right); the lower ones are for $a = 0.99$ ($Q = 0$ left and $Q = 40$ right).

5.3.3 Motion in the equatorial plane

For particles moving in the equatorial plane of a rotating black hole, the expressions for $dr/d\tau$ and $d\phi/d\tau$ can be written in the form

$$r^3 \left(\frac{dr}{d\tau} \right)^2 = \widetilde{E}^2(r^3 + a^2 r + 2Ma^2) - 4aM\widetilde{E}l_z - (r - 2M)l_z^2 - r\Delta$$

$$\tag{5.64}$$

$$\frac{d\phi}{d\tau} = \frac{(r - 2M)l_z + 2aM\widetilde{E}}{r\Delta}. \tag{5.65}$$

They are analogous to the corresponding equations for a Schwarzschild black hole. An analysis of the peculiarities of motion is performed in the same way as before by using the effective potential.

5.3.3.1 Circular orbits

The most important class of orbits is circular orbits. For given $\widetilde{E}$ and l_z, the radius r_0 of a circular orbit can be found by solving simultaneously the equations

$$\mathcal{R}(r_0) = 0 \qquad \left. \frac{d\mathcal{R}}{dr} \right|_{r_0} = 0. \tag{5.66}$$
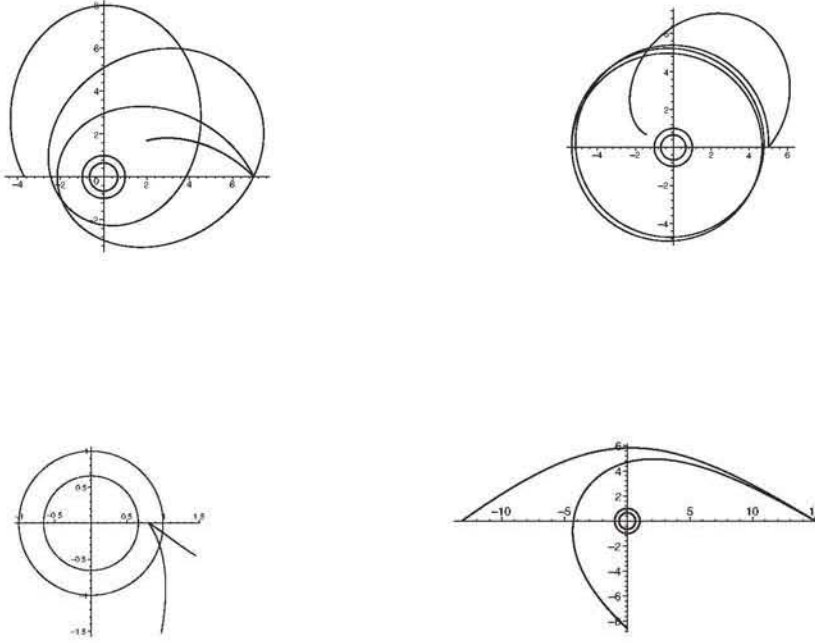
Figure 5.7. Trajectories of particles in the equatorial plane. In each case two trajectories are shown. Both trajectories have the same initial conditions. The particle is moving in the Kerr metric with $a = |a|$ and $a = -|a|$, respectively.

One can also use these equations to obtain the expressions for the specific energy $\widetilde{E}_{\text{circ}}$ and specific angular momentum l_{circ} as functions of the radius r of the circular motion [15],

$$\widetilde{E}_{\text{circ}} = \frac{r^2 - 2Mr \pm a\sqrt{Mr}}{r(r^2 - 3Mr \pm 2a\sqrt{Mr})^{1/2}} \tag{5.67}$$

$$l_{\text{circ}} = \pm \frac{\sqrt{Mr}(r^2 \mp 2a\sqrt{Mr} + a^2)}{r(r^2 - 3Mr \pm 2a\sqrt{Mr})^{1/2}}. \tag{5.68}$$

The upper signs in these and the subsequent formulas correspond to direct orbits (i.e. co-rotating with $l_z > 0$), and the lower signs correspond to retrograde orbits (counter-rotating with $l_z < 0$). We always assume that $a \geq 0$.

The coordinate angular velocity of a particle on the circular orbit is

$$\omega_{\text{circ}} = \frac{d\phi}{dt} = \frac{\pm\sqrt{Mr}}{r^2 \pm a\sqrt{Mr}}. \tag{5.69}$$
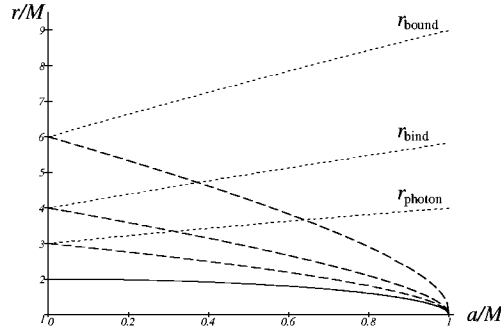
Figure 5.8. r_{photon}, r_{bind}, and r_{bound} as functions of the rotation parameter a/M. The quantities corresponding to the direct and retrograde motions are shown by dashed and dotted lines, respectively.

5.3.3.2 Last stable circular orbits

Circular orbits can exist only for those values of r for which the denominator in the expressions for $\widetilde{E}_{\text{circ}}$ and l_{circ} is real, i.e.

$$r^2 - 3Mr \pm 2a\sqrt{Mr} \geq 0. \tag{5.70}$$

The radius of the circular orbit closest to the black hole (the motion along it occurs at the speed of light) is

$$r_{\text{photon}} = 2M \left\{ 1 + \cos\left[\frac{2}{3} \arccos\left(\mp \frac{a}{M} \right) \right] \right\}. \tag{5.71}$$

This orbit is unstable. For $a = 0$, we have $r_{\text{photon}} = 3M$, while for $a = M$, we find $r_{\text{photon}} = M$ (direct motion) or $r_{\text{photon}} = 4M$ (retrograde motion).

The circular orbits with $r > r_{\text{photon}}$ and $\widetilde{E} \geq 1$ are unstable. A small perturbation directed outwards forces the particle to leave its orbit and escape to infinity on an asymptotically hyperbolic trajectory.

The radius of the unstable circular orbit, on which $\widetilde{E}_{\text{circ}} = 1$, is given by

$$r_{\text{bind}} = 2M \mp a + 2M^{1/2}(M \mp a)^{1/2}. \tag{5.72}$$

These values of the radius are the minima of periastra of all parabolic orbits. A particle in the equatorial plane, coming from infinity where its velocity is $v_\infty \ll c$, is captured if it passes the black hole closer than r_{bind}.

Finally the radius of the boundary circle separating stable circular orbits from unstable ones is given by the expression

$$r_{\text{bound}} = M\{3 + Z_2 \mp [(3 - Z_1)(3 + Z_1 + 2Z_2)]^{1/2}\} \tag{5.73}$$

Table 5.1. The radii r_{photon}, r_{bind}, and r_{bound} (in units of $r_S = 2M$) for a non-rotating ($a = 0$) and an extremely rotating ($a = M$) black hole.

Orbit	$a = 0$	$a = M$	
		$L > 0$	$L < 0$
r_{photon}	1.5	0.5	2.0
r_{bind}	2.0	0.5	2.92
r_{bound}	3.0	0.5	4.5

Table 5.2. Specific energy $\widetilde{E}$, specific binding energy $1 - \widetilde{E}$, and specific angular momentum $|l_z|/M$ of a test particle at the last stable circular orbit.

Orbit	$a = 0$	$a = M$	
		$L > 0$	$L < 0$
$\widetilde{E}$	$\sqrt{8/9}$	$\sqrt{1/3}$	$\sqrt{25/27}$
$1 - \widetilde{E}$	0.0572	0.4236	0.0377
$|l_z|/M$	$2\sqrt{3}$	$2/\sqrt{3}$	$22/3\sqrt{3}$

where

$$Z_1 = 1 + (1 - a^2/M^2)^{1/3}[(1 + a/M)^{1/3} + (1 - a/M)^{1/3}] \qquad (5.74)$$

$$Z_2 = (3a^2/M^2 + Z_1^2)^{1/2}. \qquad (5.75)$$

The quantities r_{photon}, r_{bind}, and r_{bound} as the functions of the rotation parameter a/M are shown in figure 5.8.

Table 5.1 lists r_{photon}, r_{bind}, and r_{bound} for the black hole rotating at the limiting angular velocity, $a = M$, and gives a comparison with the case of $a = 0$ (in units of $r_S = 2M$). As $a \rightarrow M$, the invariant distance from a point r to the horizon r_+,

$$\int_{r_+}^{r} \frac{r' \, dr'}{\Delta^{1/2}(r')} \qquad (5.76)$$

diverges. This does not mean that all orbits coincide in this limit and lie at the horizon, although at $L > 0$ the radii r of all three orbits tend to the same limit r_+ [15].

Finally, we will give the values of specific energy $\widetilde{E}$, specific binding energy $1 - \widetilde{E}$, and specific angular momentum $|l_z|/M$ of a test particle at the last stable circular orbit, r_{bound} (see table 5.2).

The binding energy has a maximum for an extremely rotating black hole with $a = M$. It is equal to

$$E_{\text{binding}} = (1 - 1/\sqrt{3})mc^2 \approx 0.4226\, mc^2. \tag{5.77}$$

Thus, the maximum efficiency of the energy release by matter falling into a rotating black hole is 42%. This is much higher than in a non-rotating case.

5.3.3.3 Motion with negative $\widetilde{E}$

It is easy to show that orbits with negative $\widetilde{E}$ are possible within the ergosphere for any $\theta \neq 0, \pi$. This follows from the fact that the Killing vector $\xi_{(t)}$ is spacelike inside the ergosphere. The specific energy $\widetilde{E}$ is defined as $\widetilde{E} = -u_\mu \xi^\mu_{(t)}$. Local analysis shows that for a fixed spacelike vector $\xi_{(t)}$ it is always possible to find a timelike or null vector u^μ representing the four-velocity of a particle or a photon so that $\widetilde{E}$ is negative. Orbits with $\widetilde{E} < 0$ make it possible to devise processes that extract the 'rotational energy' of the black hole. Such processes were discovered by Penrose [16].

5.3.4 Motion off the equatorial plane

We consider only a special type of motion off the equatorial plane when particles are moving quasiradially along the trajectories on which the value of the polar angle θ remains constant, $\theta = \theta_0$. For this motion

$$\Theta(\theta_0) = 0 \qquad \left.\frac{d\Theta}{d\theta}\right|_{\theta_0} = 0. \tag{5.78}$$

If we exclude trivial solutions $\theta_0 = 0$, $\theta_0 = \pi$, and $\theta_0 = \pi/2$, the relations between the integrals of motion can be written in the form

$$l_z^2 = a^2(\widetilde{E}^2 - 1)\sin^4\theta_0 \tag{5.79}$$

$$\mathcal{Q} = -a^2(\widetilde{E}^2 - 1)\cos^4\theta_0. \tag{5.80}$$

Hence, motion with constant $\theta = \theta_0$ is possible only when $\widetilde{E} > 1$ (infinite motion).

Non-relativistic particles moving at parabolic velocity ($v_\infty = 0$) and with zero angular momentum ($l_z = 0$) represent a special limiting case. Such particles fall at constant θ and are dragged into the rotation around the black hole.

Another important limiting case is the falling of ultrarelativistic particles (photons) which move at infinity at $\theta = \text{constant}$. In this limit, $\widetilde{E} \to \infty$ and $l_z \to \infty$ while their ratio $b = l_z/\widetilde{E}$ remains finite and equal to $b = a\sin^2\theta$. The null vector n^μ tangent to a null geodesic representing the motion of the in-coming photon is

$$n^\mu = \left(\frac{(r^2 + a^2)}{\Delta}, -1, 0, \frac{a}{\Delta}\right). \tag{5.81}$$

If one substitutes 1 instead of -1 into the right-hand side of this expression one obtains a congruence of outgoing photons. These two null congruences are known as the *principal null congruences* of the Kerr metric. They are *geodesic* and *shear free*. They satisfy the following relations:

$$C_{\alpha\beta\gamma[\delta}n_{\lambda]}n^{\beta}n^{\gamma} = 0 \tag{5.82}$$

where $C_{\alpha\beta\gamma\delta}$ is the Weyl tensor. Since the Kerr metric is a vacuum solution, the Weyl tensor is equal to the Riemann tensor. The principal null vectors in the Kerr geometry also obey the relation

$$\xi_{(t)\mu;\nu}n^{\nu} = \pm\frac{1}{2}\frac{\partial F}{\partial r}n_{\mu} \tag{5.83}$$

where $F = -g_{tt} = 1 - 2Mr/\Sigma$.

5.3.5   Gravitational capture

5.3.5.1   Gravitational capture of non-relativistic particles

We now consider the gravitational capture of particles by a rotating black hole (see also a review article [17]). The impact parameter $b_{\perp}$ of capturing a non-relativistic particle moving in the equatorial plane is given by the expression:

$$b_{\perp} = \pm 2M\frac{1}{v_{\infty}}\left(1 + \sqrt{1 \mp \frac{a}{M}}\right). \tag{5.84}$$

The capture cross section for particles falling perpendicularly to the rotation axis of the black hole with $a = M$ is [18]

$$\sigma_{\perp} = 14.2\pi(1/v_{\infty})^2 M^2. \tag{5.85}$$

The impact parameter of particles falling parallel to the rotation axis, $b_{\parallel}$, can be found in the following manner. Let us denote $\tilde{b}_{\parallel} = b_{\parallel}/M$, $\tilde{a} = a/M$. Then $\tilde{b}_{\parallel}$ is found as the solution of the equation

$$(1 - \tilde{a}^2)q_0^4 + 4(5\tilde{a}^2 - 4)q_0^3 - 8\tilde{a}^2(6 + \tilde{a}^2)q_0^2 - 48\tilde{a}^4 q_0 - 16\tilde{a}^6 = 0 \tag{5.86}$$

where $q_0 = v_{\infty}^2(\tilde{b}_{\parallel}^2 - \tilde{a}^2)$. If $\tilde{a} = 1$, then

$$\tilde{b}_{\parallel} = 3.85\left(\frac{1}{v_{\infty}}\right)M \qquad \sigma_{\parallel} = 14.8\pi\left(\frac{1}{v_{\infty}}\right)^2 M^2. \tag{5.87}$$

5.3.5.2   Gravitational capture of ultrarelativistic particles

Consider now ultrarelativistic particles. The impact parameters of capture, $b_{\perp}$, for the motion in the equatorial plane are given by the following formulas.

If the angular momentum is positive, then

$$\frac{b_\perp^+}{M} = 8\cos^3\left[\frac{1}{3}(\pi - \arccos\tilde{a})\right] + \tilde{a}. \tag{5.88}$$

If the angular momentum is negative, then

$$\frac{b_\perp^-}{M} = -8\cos^3\left(\frac{1}{3}\arccos|\tilde{a}|\right) + \tilde{a}. \tag{5.89}$$

In this case the cross section for $\tilde{a} = 1$ is

$$\sigma_\perp = 24.3\pi M^2. \tag{5.90}$$

For photons propagating parallel to the rotation axis of the black hole with $\tilde{a} = 1$, we have

$$\frac{b_\|}{M} = 2(1 + \sqrt{2}) \qquad \sigma_\| = 23.3\pi M^2. \tag{5.91}$$

A rotating black hole captures incident particles with a lower efficiency than a non-rotating black hole of the same mass does.

5.4 Propagation of fields in the black hole spacetime

There are many problems of black hole physics which require detailed knowledge of propagation of physical fields in the black hole geometry. In particular, they include:

- the radiation emitted by objects falling into a black hole,
- the gravitational radiation during a slightly non-spherical gravitational collapse,
- scattering and absorption of waves by a black hole,
- gravitational radiation from coalescing compact binary systems,
- analysis of stability of black hole solutions and
- quantum radiation of black holes.

For these and other relevant problems it is often sufficient to consider physical fields (including gravitational perturbations) in a linear approximation and to neglect their back reaction on the background black hole geometry.

5.4.1 Scalar massless field in the Schwarzschild metric

5.4.1.1 Field equation

The electromagnetic field and gravitational perturbations are of most interest in astrophysical applications. Both fields are massless and carry spin. For simplicity, we consider first a massless scalar field with zero spin and discuss later effects

caused by spin. Moreover, we consider the simpler case of a non-rotating black hole.

A massless scalar field evolves according to the *Klein–Gordon equation*

$$\Box\Phi := (-g)^{-1/2}\partial_\mu[(-g)^{1/2}g^{\mu\nu}\partial_\nu\Phi] = -4\pi J \tag{5.92}$$

where g is the determinant of the metric $g_{\mu\nu}$ and J a scalar charge density.

5.4.1.2 Spherical reduction

In a general spherically symmetric spacetime with metric

$$ds^2 = \gamma_{AB}\,dx^A\,dx^B + r^2\,d\Omega^2 \qquad A, B = 0, 1 \tag{5.93}$$

one can decompose a general solution into the spherical modes

$$\Phi_{\ell m} = \frac{u_\ell(t, r)}{r}Y_{\ell m}(\theta, \phi) \tag{5.94}$$

where $Y_{\ell m}(\theta, \phi)$ are the *spherical harmonics*. The functions $u_{\ell m}$ obey the two-dimensional wave equation

$$(^2\Box - V_\ell)u_{\ell m} = -4\pi j_{\ell m} \tag{5.95}$$

where $^2\Box = (-\gamma)^{-1/2}\partial_A[(-\gamma)^{1/2}\gamma^{AB}\partial_B]$ is the two-dimensional 'box'-operator for the metric γ_{AB}, and

$$V_\ell = \frac{\ell(\ell + 1)}{r^2} + \frac{^2\Box r}{r}. \tag{5.96}$$

For the Schwarzschild metric, we find

$$\frac{^2\Box r}{r} = \frac{r_S}{r^3}. \tag{5.97}$$

5.4.1.3 Radial equation and effective potential

We focus now on solutions of the homogeneous equations with $j_{\ell m} = 0$. Since the Schwarzschild geometry is static we can decompose them into monochromatic waves

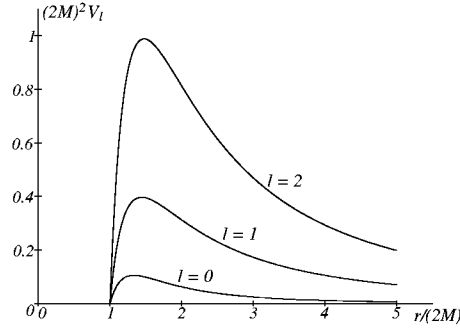$$u_{\omega\ell} = \hat{u}_\ell(r, \omega)e^{-i\omega t} \tag{5.98}$$

where the radial function $\hat{u}_\ell(r, \omega)$ is a solution of the equation

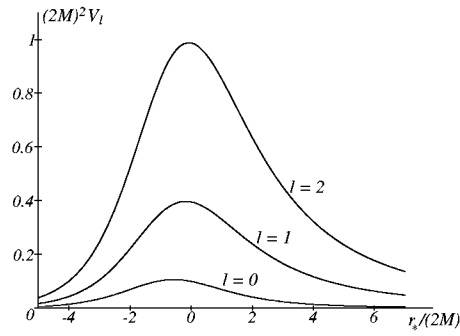$$\left[\frac{\partial^2}{\partial r_*^2} + \omega^2 - V_\ell(r)\right]\hat{u}_\ell(r, \omega) = 0. \tag{5.99}$$

Here r_* is the so-called *tortoise coordinate*

$$r_* = r + r_S\log\left(\frac{r}{r_S} - 1\right) + \text{constant} \tag{5.100}$$

Figure 5.9. The effective potential V_ℓ for $\ell = 0, 1, 2$ as a function of r.

Figure 5.10. The same potentials as functions of r_*. The constant in the definition of r_* is fixed so that $r_* = 0$ for $r = 3M$.

and $V_\ell(r)$ is the *effective potential*

$$V_\ell(r) = \left(1 - \frac{r_S}{r}\right)\left[\frac{\ell(\ell+1)}{r^2} + \frac{r_S}{r^3}\right]. \tag{5.101}$$

The effective potential for different values of ℓ is shown in figures 5.9 and 5.10. The maximum of the effective potential $V_\ell(r)$ is roughly at the location of the unstable circular photon orbit ($r = 3M$).

The form of the radial wave equation is similar to the quantum mechanical equation for one-dimensional potential scattering and hence most problems concerning perturbed black holes involve elements familiar from potential scattering in quantum mechanics. One would, for example, expect waves of short wavelength $\lambda \ll r_S$ to be easily transmitted through the barrier. Waves with $\lambda \approx r_S$ will be partly transmitted and partly reflected, and finally waves with $\lambda \gg r_S$ should be completely reflected by the black hole barrier.

5.4.1.4   Basic solutions

Let us consider two linearly independent solutions of the radial wave equations which have the following asymptotic forms

$$\hat{u}_\ell^{\text{in}}(r_*, \omega) \sim \begin{cases} e^{-i\omega r_*} & r_* \to -\infty \\ A_{\text{out}}(\omega)e^{i\omega r_*} + A_{\text{in}}(\omega)e^{-i\omega r_*} & r_* \to +\infty \end{cases} \quad (5.102)$$

$$\hat{u}_\ell^{\text{up}}(r_*, \omega) \sim \begin{cases} B_{\text{out}}(\omega)e^{i\omega r_*} + B_{\text{in}}(\omega)e^{-i\omega r_*} & r_* \to -\infty \\ e^{+i\omega r_*} & r_* \to +\infty. \end{cases} \quad (5.103)$$

They are known as IN and UP modes, respectively. The complex conjugates of these solutions are also solutions of the radial wave equation. They are known as OUT and DOWN modes, respectively.

For any two solutions u_1 and u_2 the Wronskian

$$W(u_1, u_2) = u_1 \, du_2/dr_* - u_2 \, du_1/dr_* \quad (5.104)$$

is constant. Calculating the Wronskian for solutions $\hat{u}_\ell^{\text{in}}$ and $\hat{u}_\ell^{\text{up}}$ and their complex conjugates and using their asymptotics, one obtains the following relations:

$$1 + |A_{\text{out}}|^2 = |A_{\text{in}}|^2 \quad (5.105)$$

$$1 + |B_{\text{in}}|^2 = |B_{\text{out}}|^2 \quad (5.106)$$

$$B_{\text{out}}(\omega) = A_{\text{in}}(\omega) \qquad B_{\text{in}}(\omega) = -\bar{A}_{\text{out}}(\omega) = -A_{\text{out}}(-\omega) \quad (5.107)$$

and

$$W(\hat{u}_\ell^{\text{in}}, \hat{u}_\ell^{\text{up}}) = 2i\omega A_{\text{in}}(\omega) = 2i\omega B_{\text{out}}(\omega). \quad (5.108)$$

Since $|A_{\text{in}}| \geq 1$, the solutions $\hat{u}_\ell^{\text{in}}$ and $\hat{u}_\ell^{\text{up}}$ are linearly independent.

5.4.1.5   Interpretation of basic solutions

Let us discuss the physical meaning of the basic solutions. By combining the radial solutions with $\exp(-i\omega t)$ we get functions describing wave propagation. They have simple physical interpretations. The DOWN solution satisfies the boundary condition that there is no radiation escaping to infinity. This means that exactly the right amount of radiation with just the right phase must emerge from the past horizon H^- in order to cancel any radiation that might otherwise be scattered back to infinity from a wave originally incoming from past infinity. Thus, in this solution, there is radiation coming in from infinity, radiation emerging from H^- to meet it, and radiation going down the black hole at H^+. The amplitudes of the various waves are such that down is an acceptable solution to the radial wave equation. The UP mode is defined analogously by the boundary condition that there be no incoming radiation from infinity. In a similar way the IN solution does not contain radiation outgoing from H^-, while the OUT mode has no radiation going down the black hole at H^+. The situation is presented graphically in figure 5.11.
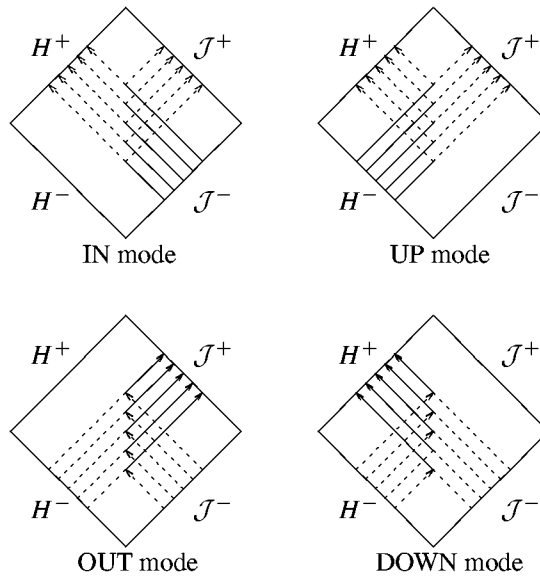
Figure 5.11. IN, UP, OUT and DOWN modes.

One can use the diagrams presented in figure 5.11 as mnemonic rules for the definition of the basic functions. The regions inside the squares represent the spacetime in the exterior of the eternal version of the black hole. The straight lines at the angle of $\pi/4$ represent null rays. Two boundaries $\mathcal{J}^+$ and $\mathcal{J}^-$ correspond to asymptotic future and past infinities. The other two boundaries H^+ and H^- are the event horizon and the past horizon, respectively. This type of diagram can be obtained by special conformal transformations that bring infinitely distant points of the spacetime to a finite distance, see chapter 1. The corresponding *Penrose–Carter conformal diagram* proved to be a very powerful tool for the study of the global structure of spacetime. Asymptotic values of massless fields in an asymptotically flat physical spacetime are related to the boundary values at the null surfaces $\mathcal{J}^+$ and $\mathcal{J}^-$, representing the so-called future and past null infinities.

5.4.2 Evolution of the scalar massless field around a non-rotating black hole

5.4.2.1 Retarded Green's function

Time evolution of the scalar massless field around a non-rotating black hole for given initial data and source j_ℓ can be easily obtained by using a *retarded Green's*

function G^{ret}. The retarded Green's function is a solution of the equation

$$\left[\frac{\partial^2}{\partial r_*^2} - \frac{\partial^2}{\partial t^2} - V_\ell(r)\right] G^{\text{ret}}(r_*, r_*', t) = \delta(t - t')\delta(r_* - r_*') \qquad (5.109)$$

which is singled out by the condition $G(r_*, r_*', t - t') = 0$ for $t < t'$.

The integral transform, reducing the problem to an ordinary differential equation,

$$\hat{G}^{\text{ret}}(r_*, r_*', \omega) = \int_{0^-}^{+\infty} G^{\text{ret}}(r_*, r_*', t)e^{i\omega t}\, dt \qquad (5.110)$$

is well defined as long as $\text{Im}\,\omega \geq 0$. In fact, $\hat{G}^{\text{ret}}(r_*, r_*', \omega)$ is a holomorphic function of $\omega = \omega_0 + i\omega_1$ for $\omega_1 > 0$. By a change $s = -i\omega$, one can show that this integral transform is nothing but the usual Laplace transform. By making the inverse Laplace transformation one obtains

$$G^{\text{ret}}(r_*, r_*', t) = \frac{1}{2\pi}\int_{-\infty+ic}^{+\infty+ic} \hat{G}^{\text{ret}}(r_*, r_*', \omega)e^{-i\omega t}\, d\omega \qquad (5.111)$$

where c is some positive number (this ensures convergence of the integral).

5.4.2.2   Green's function representation

The integral transform of the retarded Green's function obeys the equation

$$\left[\frac{d^2}{dr_*^2} + \omega^2 - V_\ell(r)\right] \hat{G}^{\text{ret}}(r_*, r_*', \omega) = \delta(r_* - r_*'). \qquad (5.112)$$

Since the retarded Green's function $G^{\text{ret}}(x, x')$ vanishes when a point x' lies to the past of x, in its decomposition there must be no waves which emerge from $\mathcal{J}^-$ and H^-. Thus $\hat{G}^{\text{ret}}(r_*, r_*', \omega)$ can be written as

$$\hat{G}^{\text{ret}}(r_*, r_*', \omega) = -\frac{1}{2i\omega A_{\text{in}}(\omega)}\begin{cases} \hat{u}_\ell^{\text{in}}(r_*, \omega)\hat{u}_\ell^{\text{up}}(r_*', \omega) & r_* < r_*' \\ \hat{u}_\ell^{\text{in}}(r_*', \omega)\hat{u}_\ell^{\text{up}}(r_*, \omega) & r_* > r_*'. \end{cases} \qquad (5.113)$$

The factor containing A_{in} is the Wronskian.

5.4.2.3   Analytical properties

In order to infer the behavior of the Green's function in different time intervals it is convenient to deform the contour of integration in the complex ω-plane. For this purpose we need to know the analytic properties of $\hat{u}_\ell^{\text{in}}$ and $\hat{u}_\ell^{\text{up}}$, not only in the upper half-plane where they are a holomorphic functions of ω, but also in the lower half-plane. A detailed analysis of this problem can be found in [19]. Here we just describe the most important results.
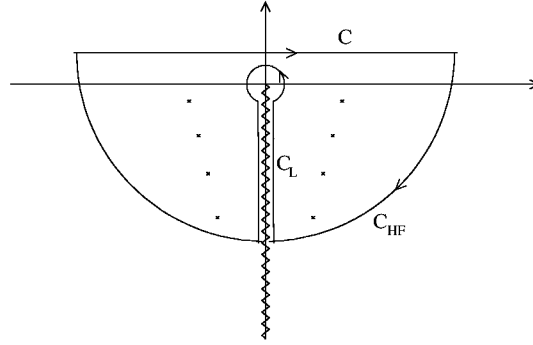
Figure 5.12. Integration contours in the complex ω-plane. The crosses represent the first few quasinormal modes. The necessary branch cut is taken along the negative imaginary axis.

The analysis shows that the Wronskian $W(\text{IN,UP}) = 2i\omega A_{\text{in}}(\omega)$ has isolated zeros there. This leads to poles of the Green's function $\hat{G}^{\text{ret}}(r_*, r'_*, \omega)$. These singularities correspond directly to the so called *quasinormal modes* of the black hole. It is straightforward to show that the poles are symmetrically distributed with respect to the imaginary ω-axis; if ω_n corresponds to $A_{\text{in}} = 0$, then $-\bar{\omega}_n$ must also do so (see figure 5.12).

In the upper half of the complex ω-plane, the solutions, which are bound at either end, must behave like

$$
\begin{aligned}
\hat{u}_{\ell}^{\text{in}}(r_*, \omega) &\sim e^{-i\omega r_*} & \text{for } r_* \to -\infty \\
\hat{u}_{\ell}^{\text{up}}(r_*, \omega) &\sim e^{+i\omega r_*} & \text{for } r_* \to \infty.
\end{aligned}
\tag{5.114}
$$

Their analytical continuations into the lower half-plane will show the same behavior. Hence, the Green's function always satisfies 'future outgoing' conditions. This Green's function propagates waves emitted by the source to H^+ and $\mathcal{J}^+$. It is therefore clear that the solutions corresponding to the quasinormal modes are regular both at H^+ and $\mathcal{J}^+$. But it also follows that they will diverge at H^- and $\mathcal{J}^-$.

Careful analysis shows that it is necessary to introduce a branch cut in order to make $\hat{u}_{\ell}^{\text{up}}$ a single-valued function [20]. This cut is usually placed along the negative imaginary axis, as in figure 5.12.

Given this information, the radiation produced in response to a perturbation of the black hole can be divided into three components, in accordance with the contributions from different parts of the deformed contour in the lower half of the ω-plane:

(i)   radiation emitted directly by the source,

(ii) exponentially damped quasinormal-mode oscillations (contribution of the poles of the Green's function) and

(iii) a power-law tail (contribution of the branch cut integral).

5.4.2.4   Quasinormal modes

Scattering resonances (which are the quantum analogues to quasinormal modes) arise for energies close to the top of a potential barrier. In the black hole case, this immediately leads to the approximation [21–23]

$$\mathrm{Re}\,\omega_0 \approx \sqrt{V_\ell^{\max}} \approx \frac{1}{3\sqrt{3}M}\left(\ell + \frac{1}{2}\right). \tag{5.115}$$

This approximation for the fundamental mode is poor for low ℓ (the error is something like 30% for $\ell = 2$) but it rapidly gets accurate as ℓ increases.

For the imaginary part of the frequency—the lifetime of the resonance—the curvature of the potential contains the relevant information [21]. One finds that

$$\mathrm{Im}\,\omega_0 \approx -\frac{1}{2}\left|\frac{1}{2V_\ell}\frac{d^2V_\ell}{dr_*^2}\right|_{r=r_{\max}}^{1/2} \approx -\frac{\sqrt{3}}{18M} \tag{5.116}$$

which is accurate to within 10% for the fundamental mode.

Interestingly, similar approximations follow from a different approach. Consider a congruence of null rays circling the black hole in the unstable photon orbit at $r = 3M$. The fundamental mode frequency then follows if the beam contains ℓ cycles [24]. The damping rate of the mode can be inferred from the decay rate of the congruence if the null orbit is slightly perturbed [25].

It is interesting to compare a black hole to other resonant systems in nature. If we define a quality factor in analogy with the standard harmonic oscillator,

$$Q \approx \frac{1}{2}\left|\frac{\mathrm{Re}\,\omega_n}{\mathrm{Im}\,\omega_n}\right| \tag{5.117}$$

the quasinormal-mode approximations given here lead to $Q \approx \ell$. This should be compared to the typical value for an atom: $Q \sim 10^6$. The Schwarzschild black hole is thus a very poor oscillator.

5.4.2.5   Late-time behavior

The power-law tail is associated with the branch-cut integral along the negative imaginary axis in the complex ω-plane. The main contribution gives the $|\omega M| \ll 1$ part of that integral. A branch-cut contribution to the Green's function is [19,20]

$$G_{\mathrm{B}}^{\mathrm{ret}}(r_*, r_*', t) = (-1)^{\ell+1}\frac{(2\ell + 2)!}{[(2\ell + 1)!!]^2}\frac{4M(r_*r_*')^{\ell+1}}{t^{2\ell+3}}. \tag{5.118}$$

This result implies that if the source of radiation falls down beyond the potential barrier the damping of its radiation that is seen by a distant observer is not purely exponential. The late-time behavior of the field is

$$\Phi \sim t^{-(2\ell+3)}. \tag{5.119}$$

This power-law behavior is connected with the scattering of emitted radiation by the 'tail' of the potential barrier (by the spacetime curvature).

Price [26] found that backscattering by the asymptotic 'tail' of the potential gives rise to a power-law fall-off at late times. Price put his conclusions in the following succinct form: 'Anything that can be radiated will be radiated.' Consequently, a black hole gets rid of all bumps after it is formed by a non-spherical collapsing star.

5.4.3 Wave fields in the Kerr metric

5.4.3.1 *Electromagnetic waves and gravitational perturbations in the Kerr geometry*

The scalar massless field we have considered is a toy model. For astrophysical applications it is important to understand the behavior of electromagnetic waves and (in view of coming gravitational wave experiments) gravitational perturbations. Aspects of the problem connected with the black hole rotation also might be important. Fortunately, as demonstrated by Teukolsky [27], the initial set of equations, which describes electromagnetic waves or gravitational perturbations in the Kerr metric, can be reduced to a form which allows *decoupling*. Moreover, the resulting decoupled equations allow *separation of variables*. We describe here only the scheme and the main result, omitting details and long calculations.

The homogeneous equations describing electromagnetic waves or gravitational perturbations propagating in an external gravitational field are of the form

$$\mathcal{E}^{AB}\varphi_B = 0 \tag{5.120}$$

where $\mathcal{E}^{AB}$ is a covariant differential operator, and A and B represent collective tensorial indices. For the electromagnetic field

$$\mathcal{E}^{\mu\nu}A_\nu = \nabla^\nu\nabla_\nu A^\mu - \nabla^\nu\nabla^\mu A_\nu \tag{5.121}$$

and for gravitational perturbations

$$\mathcal{E}^{\mu\nu\alpha\beta}h_{\alpha\beta} = -\nabla^\mu\nabla^\nu h^\alpha{}_\alpha - \nabla^\alpha\nabla_\alpha h^{\mu\nu} + \nabla^\alpha\nabla^\nu h_\alpha{}^\mu + \nabla^\alpha\nabla^\mu h_\alpha{}^\nu$$
$$+ g^{\mu\nu}(\nabla^\alpha\nabla_\alpha h^\beta{}_\beta - \nabla^\alpha\nabla^\beta h_{\alpha\beta}). \tag{5.122}$$

5.4.3.2 Field equation decoupling. Teukolsky equation

In the Kerr metric these equations (as well as the equations for other massless fields with spin s) can be decoupled. This means that there exist three operators (we denote them $_s\tau_A$, $_s\Pi^B$, and $_s\widetilde{\square}$) such that the following relation is valid:

$$_s\tau_A \mathcal{E}^{AB} = {}_s\widetilde{\square}\, _s\Pi^B. \tag{5.123}$$

Here $s = 0, \pm 1/2, \pm 1, \pm 3/2, \pm 2$, and $|s|$ is the spin of the field φ_A. For electromagnetic waves $|s| = 1$ and for gravitational perturbations $|s| = 2$. Relation (5.123) shows that the scalar $_s\psi = {}_s\Pi^B\varphi_B$, constructed for any solution of the equation $\mathcal{E}^{AB}\varphi_B = 0$, obeys the scalar decoupled equation

$$_s\widetilde{\square}\, _s\psi = 0. \tag{5.124}$$

Usually the covariant operator $_s\widetilde{\square}$ is presented in the form

$$_s\widetilde{\square} = \Sigma^{-1}\, _s\square. \tag{5.125}$$

We recall that in the Kerr metric $\sqrt{-g} = \Sigma \sin\theta$. The scalar second-order differential operator $_s\square$ was introduced by Teukolsky [27]. Its explicit form depends on the choice of coordinates and complex null tetrads. In Boyer–Lindquist coordinates and for the so-called *Kinnersley tetrad*, the operator $_s\square$ is specified by

$$\begin{aligned}
_s\square\, _s\psi = {} & \left[\frac{(r^2 + a^2)^2}{\Delta} - a^2\sin^2\theta\right]\frac{\partial_s^2\psi}{\partial t^2} + \frac{4aMr}{\Delta}\frac{\partial_s^2\psi}{\partial t\partial\phi} \\
& + \left(\frac{a^2}{\Delta} - \frac{1}{\sin^2\theta}\right)\frac{\partial_s^2\psi}{\partial\phi^2} - \Delta^{-s}\frac{\partial}{\partial r}\left(\Delta^{s+1}\frac{\partial_s\psi}{\partial r}\right) \\
& - \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial_s\psi}{\partial\theta}\right) - 2s\left[\frac{a(r - M)}{\Delta} + \frac{i\cos\theta}{\sin^2\theta}\right]\frac{\partial_s\psi}{\partial\phi} \\
& - 2s\left(\frac{M(r^2 - a^2)}{\Delta} - r - ia\cos\theta\right)\frac{\partial_s\psi}{\partial t} + (s^2\cot^2\theta - s)_s\psi = 0.
\end{aligned} \tag{5.126}$$

This is the so-called *Teukolsky equation*.

5.4.3.3 Field restoration from solutions of the decoupled equations

A solution of the tensor field equation can be constructed from the solutions of the related Teukolsky equation. This was demonstrated by Cohen and Kegels [28] for the electromagnetic field and by Chrzanowski [29] for gravitational perturbations.

Wald [30] gave a simple proof of a general result which shows how to construct a field once one has succeeded in deriving a decoupled equation.

In order to describe the idea of this proof, let us introduce a scalar product of two (generally complex) tensor fields ψ_A and φ^A

$$(\psi, \varphi) \equiv \int \sqrt{-g}\, d^4x\, \overline{\psi}_A \varphi^A. \tag{5.127}$$

The action $W[\varphi]$ for a real tensor field φ obeying the field equation can then be written in the compact form

$$W[\varphi] = \tfrac{1}{2}(\varphi, \mathcal{E}\varphi). \tag{5.128}$$

Further denote by $Q^\star$ an operator which is conjugated to an operator Q with respect to the scalar product

$$(\psi, Q\varphi) = (Q^\star\psi, \varphi). \tag{5.129}$$

The operator $\mathcal{E}$ is then self-conjugated $\mathcal{E}^\star = \mathcal{E}$. By using this property we can rewrite the equation obtained by conjugation in the form

$$\mathcal{E}^{AB}\, {}_s\tau_A^\star = {}_s\Pi^{\star B}\, {}_s\widetilde{\square}^\star. \tag{5.130}$$

It can be shown that

$${}_s\widetilde{\square}^\star = \Sigma^{-1}\, {}_{-s}\overline{\square}. \tag{5.131}$$

This relation shows that for any solution ${}_s\Xi$ of the scalar equation

$${}_s\widetilde{\square}^\star\, {}_s\Xi = 0 \tag{5.132}$$

the tensor function

$${}_s\varphi_A = {}_s\tau_A^\star\, {}_s\Xi \tag{5.133}$$

is a solution to the field equation. Moreover, it can be shown that all such solutions (up to possible gauge transformation) can be represented in this form. Thus, solutions to the scalar decoupled equation provide complete information about the perturbing field. For more details see [29–36].

5.4.3.4 Separation of variables, spin-weighted spheroidal harmonics

The coefficients of the Teukolsky equation do not depend on t and ϕ. Furthermore, the existence of the Killing tensor in Kerr spacetime yields an additional symmetry of the Teukolsky equation which makes it possible to solve it by separation of variables and to write a solution in terms of the modes

$${}_sR_{\ell m}(r, \omega)\, {}_sZ_{\ell m}^\omega(\theta, \phi)e^{-i\omega t} \tag{5.134}$$

where $_sZ_{\ell m}^{\omega}(\theta, \phi)$ are the *spin-weighted spheroidal harmonics*

$$_sZ_{\ell m}^{\omega}(\theta, \phi) = (2\pi)^{-1/2}\, _sS_{\ell m}^{\omega}(\theta)e^{im\phi}. \qquad (5.135)$$

The angular problem reduces to one of solving

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dS}{d\theta}\right) + \left[a^2\omega^2\cos^2\theta - \frac{m^2}{\sin^2\theta} - 2a\omega s\cos\theta\right.$$

$$\left. -\frac{2ms\cos\theta}{\sin^2\theta} - s^2\cot^2\theta + E - s^2\right]S = 0 \qquad (5.136)$$

where the functions $_sS_{\ell m}(\theta)$ are regular on the interval $[0, \pi]$.

The required functions $_sS_{\ell m}(\theta)$ thus essentially follow from a Sturm–Liouville eigenvalue problem for the separation constant E. Boundary conditions of regularity should be imposed at both $\theta = 0$ and π. According to Sturm–Liouville theory, the eigenfunctions form a complete, orthogonal set on the interval $0 \le \theta \le \pi$ for each combination of s, $a\omega$ and m. This infinite set of eigenfunctions is enumerated by ℓ [31, 37]. For $a = 0$ and $s = 0$ functions $_sS_{\ell m}(\theta)$ coincide with the associated Legendre polynomials $P_\ell^m(\cos\theta)$.

5.4.3.5   The radial equation

The radial functions $_sR_{\ell m}$ obey a second-order differential equation [27]

$$\Delta^{-s}\frac{d}{dr}\left(\Delta^{s+1}\frac{dR}{dr}\right) + \left[\frac{K^2 - 2is(r - M)K}{\Delta} + 4ir\omega s - \lambda\right]R = 0 \qquad (5.137)$$

where

$$K \equiv (r^2 + a^2)\omega - am \qquad \lambda \equiv E - 2am\omega + a^2\omega^2 - s(s + 1). \qquad (5.138)$$

An important property of the radial Teukolsky equation is that the two solutions $_sR_{\ell m}$ and $_{-s}R_{\ell m}$ are related [3].

Introducing a new dependent variable

$$_s\chi_{\ell m} = (r^2 + a^2)^{1/2}\Delta^{s/2}\,_sR_{\ell m} \qquad (5.139)$$

the radial equation can be written in the form

$$\frac{d_s\chi_{\ell m}}{dr_*^2} + _sV_{\ell m s}\chi_{\ell m} = 0 \qquad (5.140)$$

which is similar to the form of the radial equation for a scalar massless field in the Schwarzschild geometry. Nevertheless there are important differences. The

effective potential $_sV_{\ell m}$ is

$$_sV_{\ell m}(r, \omega) = \frac{K^2 - 2is(r - M)K + \Delta(4i\omega sr - \lambda)}{(r^2 + a^2)^2} - G^2 - \frac{dG}{dr_*} \quad (5.141)$$

$$G = \frac{r\Delta}{(r^2 + a^2)^2} + \frac{s(r - M)}{r^2 + a^2} \quad (5.142)$$

and r_* is the 'tortoise' coordinate defined by

$$dr_* = \frac{r^2 + a^2}{\Delta} dr. \quad (5.143)$$

This potential is *complex* and it depends on the frequency ω.

In the asymptotic regions ($r_* \to \pm\infty$) the potential $_sV_{\ell m}$ takes the form

$$_sV_{\ell m}(r, \omega) = \begin{cases} \omega(\omega + 2is/r) & r_* \to \infty \\ \varpi_s^2 & r_* \to -\infty. \end{cases} \quad (5.144)$$

Here

$$\varpi_s = \varpi - is\kappa \qquad \varpi = \omega - m\Omega^H \quad (5.145)$$

and $\kappa = (r_+ - M)/(r_+^2 + a^2)$ is the *surface gravity* of the Kerr black hole. Two linearly independent solutions have the asymptotic behavior $\sim r^{\mp s} \exp(\pm i\omega r_*)$ at $r \to \infty$ and $\sim \Delta^{\pm s/2} \exp(\pm i\varpi r_*)$ at $r \to r_+$.

5.4.3.6 Modes

In the same manner as it was done for a massless scalar field in the Schwarzschild geometry, solutions to the equations describing a free massless field in the exterior of a Kerr black hole can be specified by prescribing their asymptotics at null infinity and the horizon. It is possible to introduce four sets of solutions called IN, UP, OUT, and DOWN modes, correspondingly in the same manner as it was done for the Schwarzschild spacetime.

Each of the modes is a solution characterized by the set $\{\ell m\omega P\}$ of quantum numbers, where the spirality $P = \pm 1$. These modes are singled out by the following requirements: IN modes vanish at H^- and have non-vanishing asymptotics at $\mathcal{J}^-$. UP modes vanish at $\mathcal{J}^-$ and are non-vanishing on the past horizon H^-. OUT modes vanish at H^+ but not at $\mathcal{J}^+$, and finally the DOWN modes vanish at $\mathcal{J}^+$ and are non-vanishing on the past horizon H^+. Any two of these four solutions can be used as a complete set in the space of solutions.

5.4.4 Effects connected with black hole rotation

5.4.4.1 Wave evolution and quasinormal modes in the Kerr spacetime

Using mode expansion one can construct a retarded Green's function as it was done for the Schwarzschild spacetime. By studying the analytical properties of

the modes and the Green's function as functions of the complex frequency ω one can demonstrate that in the general case of a rotating black hole the time evolution of radiation from a source of the perturbation is qualitatively the same as for the Schwarschild black hole. Namely, the emitted radiation consists of the following three components:

(i)   an initial wave burst that contains radiation emitted directly by the source of the perturbation,

(ii) exponentially damped 'ringing' at frequencies that do not depend on the source of the perturbation at all and

(iii) a power-law 'tail' that arises because of backscattering by the long-range gravitational field.

Quantitive differences which exist between non-rotating and rotating black hole cases are of the most interest since they, in principle, might allow an observer receiving radiation from a black hole to determine its angular velocity. Let us discuss first quasinormal modes in the Kerr spacetime.

When the black hole has nonzero angular momentum, a, the azimuthal degeneracy is split. For a multipole ℓ there are consequently $2\ell + 1$ distinct modes that approach each Schwarzschild mode in the limit $a \to 0$. These modes correspond to different values of m, where $-\ell \le m \le \ell$.

Quasinormal modes for Kerr black holes were first calculated by Detweiler [38, 39]. In the limit of the extremal black hole $(a \to M)$, complex frequencies of quasinormal modes possess the following properties:

for $m = \ell$

$$\left\{ \begin{array}{l} \text{Im}\,\omega_n \text{ is almost constant} \\ \text{Re}\,\omega_n \text{ increases monotonically} \end{array} \right\} \quad \text{as } a \to M$$

and for $m = -\ell$

$$\left\{ \begin{array}{l} \text{Im}\,\omega_n \to 0 \\ \text{Re}\,\omega_n \to -m/2 \end{array} \right\} \quad \text{as } a \to M.$$

It is interesting that some quasinormal modes become very long lived for rapidly rotating black holes. This could potentially be of great importance for gravitational-wave detection.

5.4.4.2 Gravitational radiation from a particle plunging into the black hole

In general, the equation governing a black hole perturbation is not homogeneous. One must typically also include a source term appropriate for the physical situation under consideration. Perhaps the simplest relevant problem is that of a test particle moving in the gravitational field of a black hole. When the mass m of the particle is sufficiently small compared to that of the black hole $(m \ll M)$, the problem can be viewed as a perturbation problem. The radiation emitted by a test particle of mass m which falls radially into a black hole is one of the astrophysical applications of the perturbation equations.

Simple dimensional arguments show that the total energy ΔE, emitted by the particle of mass m plunging into the black hole of mass M, is proportional to m^2/M. When the black hole is rotating, ΔE is an asymmetric function of $\tilde{L}$. As shown in [40,41], it has a minimum at the negative value of $\tilde{L}$. This can be understood in the following way. Positive values of $\tilde{L}$ correspond to a particle that corotates with the black hole, whereas negative values are for counter-rotation. When a particle that was initially counter-rotating reaches the vicinity of the black hole, it will be slowed down because of frame-dragging. Thus, fewer gravitational waves are radiated. Similarly, an initially corotating particle is speeded up, and the number of gravitational waves that emerges increases.

5.4.4.3 *Superradiant scattering*

For wave scattering by an absorbing non-rotating body, the amplitude of the reflected wave is always less than the amplitude of the infalling wave. In the presence of a ergosphere, that is, the region around a rotating black hole where $\xi_{(t)}^2 > 1$, some of the impinging waves can be amplified. This effect is known as *superradiance* [42–44]. The condition for superradiant modes is

$$\omega < m\Omega^H = \frac{ma}{2Mr_+}.\tag{5.146}$$

The maximum amplification of an incoming wave is 0.3% for scalar waves, 4.4% for electromagnetic waves, and an impressive 138% for gravitational waves [45,46].

5.5 Black hole electrodynamics

5.5.1 Introduction

Black hole electrodynamics is defined as the theory of electrodynamic processes that can occur *outside* the event horizon, accessible to observation by distant observers. At first glance, black hole electrodynamics is quite trivial. Indeed, the electromagnetic field of a stationary black hole (of a given mass M) is determined unambiguously by its electric charge Q and rotation parameter a. If the charged black hole does not rotate, its electromagnetic field reduces to the radial electric field of the charge Q and is static. Any multipoles higher than the monopole are absent.

A charged rotating black hole induces a magnetic field and distorts the geometry of space and generates higher-order electric and magnetic moments. However, these higher-order moments are determined unambiguously by the quantities M, a, and Q. These moments are not independent, as one would find in the case of ordinary bodies.

In astrophysics, the electric charge of a black hole cannot be high. The magnetic field must also be very weak: the dipole magnetic moment of a black

hole is $\mu^* = Qa$. There can be no other stationary electromagnetic field inherent to a black hole. In this sense, the electrodynamics of, say, radio pulsars possessing a gigantic 'frozen-in' magnetic field of about 10^{12} G is definitely much richer than that of the intrinsic fields of black holes.

However, if a black hole is placed in an *external* electromagnetic field and if charged particles are present in its surroundings, the situation changes dramatically and complex electrodynamics does appear. It is this aspect that we mean when black hole electrodynamics is discussed.

The case which is important for astrophysical applications is that of external magnetic (not electric) fields and rarefied plasma in which a black hole is embedded. In this system a regular magnetic field arises, for example, as it gets cleansed of magnetic loops which fall into a black hole. A regular magnetic field can also be generated in an accretion disk by the dynamo action.

In order to study the interaction of a black hole with its surrounding fields, we use the field equations and 'total absorption' boundary conditions at the surface of the black hole. The latter boundary conditions reflect the fact that the event horizon of a black hole is a null surface (at least at its regular points). Because of this property, the black hole horizon plays the role of a one-way membrane. Technically, this type of boundary condition which implies that the black hole interior cannot affect processes outside the horizon is quite simple, but it makes black holes different from usual astrophysical objects, which are bodies with a (timelike) boundary. It helps a lot, especially concerning our intuition, to develop a formalism in which black holes are more similar to ordinary physical objects. We describe briefly such an approach known as the 'membrane paradigm' and some important results of black hole electrodynamics.

5.5.2 Electrodynamics in a homogeneous gravitational field

5.5.2.1 *Electrodynamics in the uniformly accelerated frame*

In order to study the properties of the electromagnetic field in the black hole vicinity it is instructive to neglect effects connected with curvature at first. Since the curvature near the horizon is $\sim M^{-2}$, where M is the black hole mass, this limit corresponds formally to the case of $M \to \infty$. For an observer at rest near the horizon of such a black hole, the gravitational field is almost homogeneous. According to the equivalence principle, physical laws in the reference frame of the observer mentioned before are identical to the laws in flat spacetime considered in a uniformly accelerated reference frame. Let us discuss now a formulation of the standard electrodynamics in a uniformly accelerated frame.

Consider an accelerated observer in flat spacetime in Rindler coordinates,

$$ds^2 = -dT^2 + dZ^2 + dX^2 + dY^2 \tag{5.147}$$

$$T = z\sinh(w\tau) \qquad Z = z\cosh(w\tau) \qquad X = x \qquad Y = y. \tag{5.148}$$

The accelerated observer is located at $z = w^{-1}$, and w and τ are the four-
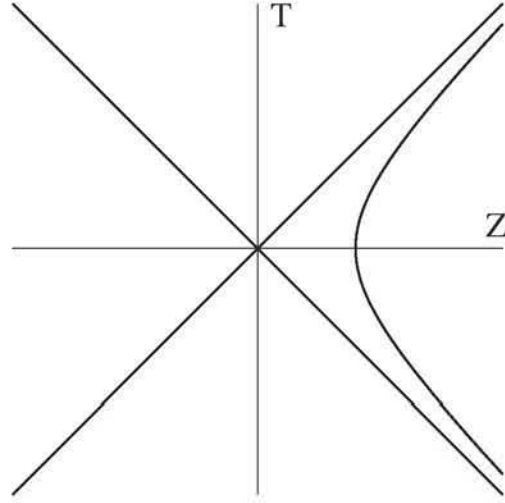
Figure 5.13. Rindler spacetime.

acceleration and proper time of the observer, respectively. The coordinates (τ, x, y, z) cover the right wedge $X > |T|$ of the Minkowski spacetime. The null surface $U = T - X = 0$ plays the role of the event horizon, since the observer cannot get any information from the region $U > 0$ lying beyond it.

The electrodynamics in the accelerated frame is described by the standard Maxwell equations

$$F^{\mu\nu}{}_{;\nu} = 4\pi J^{\mu} \qquad F_{[\mu\nu;\alpha]} = 0. \tag{5.149}$$

Since nothing that happens beyond the horizon affects the information available to the observer, let us simply assume that $F_{\mu\nu}$ vanishes in this region. Substituting the ansatz $\tilde{F}_{\mu\nu} = F_{\mu\nu}\Theta(U)$ into the Maxwell equations one gets

$$\tilde{F}^{\mu\nu}{}_{;\nu} = 4\pi(\tilde{J}^{\mu} + j^{\mu}) \qquad \tilde{F}_{[\mu\nu;\alpha]} + F_{[\mu\nu}U_{,\alpha]}\delta(U) = 0. \tag{5.150}$$

Here $\tilde{J}^{\mu} = J^{\mu}\Theta(U)$ and

$$j^{\mu} = \frac{1}{4\pi}F^{\mu\nu}U_{,\nu}\delta(U). \tag{5.151}$$

This equation shows that in order to have a vanishing field beyond the horizon there must be an additional surface current j^{μ}. We call this current *fictitious*. We introduce it only in order to mimic the correct electromagnetic field outside the horizon. Certainly, any observer crossing the horizon will not see any currents until there are real charged particles crossing the horizon. Since $j^{\mu}U_{,\mu} = 0$ this

current is propagating along the horizon. One also has

$$j^{\mu}{}_{;\mu} = -J^{\mu}U_{,\mu}.$$ (5.152)

Thus, the fictitious current is conserved until real charged particles cross the horizon.

5.5.2.2   Fictitious horizon currents for an accelerated charge

Let us consider a simple example. Suppose a pointlike electric charge q is moving with a constant acceleration w in the Z-direction. Using a standard solution of this problem in the form of the Lienard–Wiechert potentials, one can obtain the non-vanishing components of the field strength:

$$\tilde{F}_{UV} = \frac{2q}{w^2}\frac{(\rho^2 + \eta + w^{-2})}{S^3}$$ (5.153)

$$\tilde{F}_{U\rho} = -\frac{4q\rho V}{w^2 S^3}\qquad \tilde{F}_{V\rho} = \frac{4q\rho U}{w^2 S^3}$$ (5.154)

where

$$S = [(\rho^2 - \eta + w^{-2})^2 + 4\eta w^{-2}]^{1/2}$$ (5.155)

$$V = T - Z\qquad \eta = UV\qquad \rho^2 = X^2 + Y^2 = x^2 + y^2.$$ (5.156)

Simple calculations show that at the surface $U = 0$ there is only one non-vanishing component of the fictitious current

$$j^V = -\frac{2q}{\pi w^2}\frac{1}{(\rho^2 + w^{-2})^2}.$$ (5.157)

Thus, one has an axisymmetric distribution of a negative fictitious charge density

$$\sigma^H = j^T = \frac{1}{2}j^V = -\frac{q}{\pi w^2}\frac{1}{(\rho^2 + w^{-2})^2}.$$ (5.158)

We took into account that $T = V/2$ at the surface $U = 0$. It is easy to check that the integral of σ over the total X-Y plane is $-q$.

5.5.2.3   Electric charge in the homogeneous gravitational field

According to the equivalence principle physical laws in the static homogeneous gravitational field are equivalent to the laws found in a uniformly accelerated frame in flat spacetime. For our concrete case, this means the following. Let us write the flat metric in Rindler coordinates

$$ds^2 = -z^2 w^2\, d\tau^2 + dz^2 + dx^2 + dy^2.$$ (5.159)
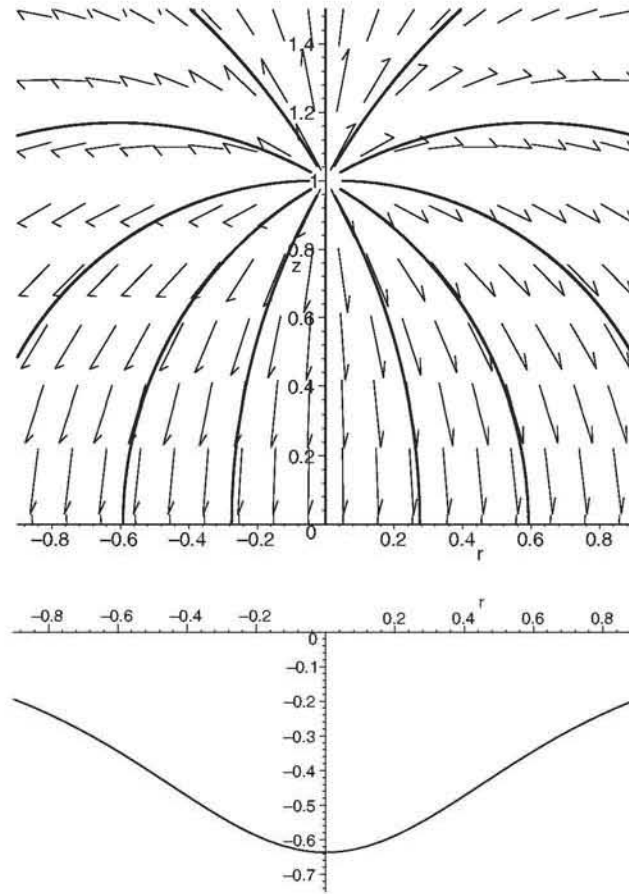
Figure 5.14. Electric field of a pointlike charge in a homogeneous gravitational field and the corresponding fictitious charge density on the horizon.

A pointlike electric charge q is at rest in this gravitational field at the point $(w^{-1}, 0, 0)$ and its electric potential is

$$A_\mu = \delta_\mu^\tau A_\tau \qquad A_\tau = -qw \frac{\rho^2 + z^2 + w^{-2}}{\sqrt{(\rho^2 + z^2 + w^{-2})^2 - 4z^2 w^{-2}}}. \qquad (5.160)$$

This is a solution of Maxwell's equations in the Rindler metric:

$$E^i{}_{,i} = 4\pi \rho_{el} \qquad i = 1, 2, 3 \qquad (5.161)$$

$E^i = (wz)^{-1} A_{\tau,i}$ is the (three-dimensional) electric field and $\rho_{el} = wzJ^\tau$.

Integral lines of the vector field E^i are shown in figure 5.14. These lines are orthogonal to the horizon. The induced negative fictitious charge density on the horizon is equal to $\sigma^H = E_z(z = 0)/4\pi$ and it coincides with the expression given in the previous section. This relation can be considered as a boundary condition at the horizon. It resembles the boundary condition at a conducting surface. In the next section we discuss this analogy in more detail.

Let us assume now that the electric charge q slowly moves away from the horizon and then returns to its initial position. Let $v = \tau + \ln(wz)$ be an advanced time. We can rewrite the conservation law for the fictitious currents on the horizon as

$$\partial_v \sigma^H + \partial_\perp j^\perp = 0. \tag{5.162}$$

This equation can be integrated for an arbitrary function $w(\tau)$ with the following result

$$j^\rho = \frac{\dot{w}}{w}\sigma^H \qquad j^\phi = 0. \tag{5.163}$$

5.5.3 Membrane interpretation

5.5.3.1 *Maxwell's equations in (3 + 1)-form*

Following [5] we define a set of observers in the Kerr spacetime which have the four-velocity u^μ such that

$$u_\mu = -\alpha\delta^t_\mu \tag{5.164}$$

where $\alpha := (\Sigma\Delta/A)^{1/2}$ is the *lapse function*. Since $u_\mu \xi^\mu_{(\phi)} = 0$, this family of observers has *zero angular momentum*. By means of the projector $h_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu$ we can obtain spatial vectors and tensors in the reference frame of the observer.

We introduce the following notation for the so defined electrodynamical quantities measured by locally non-rotating observers: **E** is the electric field strength, **B** is the magnetic field strength, ρ_e is the electric charge density, and **j** is the electric current density. Denote by ϖ the norm of a Killing vector $\xi^\mu_{(\phi)}$ reflecting the axial symmetry of spacetime,

$$\varpi \equiv \sqrt{g_{\phi\phi}} = \sqrt{\frac{A}{\Sigma}}\sin\theta. \tag{5.165}$$

We denote by $\mathbf{e}_{\hat{\phi}}$ a three-dimensional unit vector in the direction of the Killing vector $\xi^\mu_{(\phi)}$. By using this notation, Maxwell's equations can be written in the following form:

$$\nabla\mathbf{E} = 4\pi\rho_e \tag{5.166}$$

$$\nabla\mathbf{B} = 0 \tag{5.167}$$

$$\nabla \times (\alpha\mathbf{B}) = \frac{4\pi\alpha\mathbf{j}}{c} + \frac{1}{c}[\dot{\mathbf{E}} + \mathcal{L}_\beta\mathbf{E}] \tag{5.168}$$

$$\nabla \times (\alpha \mathbf{E}) = -\frac{1}{c}[\dot{\mathbf{B}} + \mathcal{L}_{\beta}\mathbf{B}].$$ (5.169)

Here

$$\boldsymbol{\beta} = \omega\varpi\mathbf{e}_{\phi}$$ (5.170)

and

$$\omega = \frac{g_{t\phi}}{g_{\phi\phi}} = \frac{2Mar}{(r+a)^2 - \Delta a^2 \sin^2\theta}$$ (5.171)

is the angular velocity of rotation (with respect to t) of locally non-rotating observers. The notation $\mathcal{L}_{\beta}\mathbf{E}$ is used for the Lie derivative of a vector $\mathbf{E}$ along $\boldsymbol{\beta}$

$$\mathcal{L}_{\beta}\mathbf{E} := (\boldsymbol{\beta}\nabla)\mathbf{E} - (\mathbf{E}\nabla)\boldsymbol{\beta}.$$ (5.172)

This Lie derivative describes how the vector $\mathbf{E}$ varies with respect to the field $\boldsymbol{\beta}$. $\mathcal{L}_{\beta}\mathbf{E}$ vanishes when the origin and the end of the vector $\mathbf{E}$ are 'glued' under a displacement by $\boldsymbol{\beta}\,d\phi$. A dot denotes differentiation with respect to t and ∇ is the three-dimensional (covariant) gradient operator in the curved 'absolute' space with metric $h_{\mu\nu}$.

The equations (5.166), (5.167) have a familiar form, whereas (5.168), (5.169) are slightly unusual. The following differences are evident. The function α appears because the physical time flows differently at different points of space while the equations are written in terms of the global 'time' t (recall that the acceleration of free fall, $\mathbf{a}$, is related to α in the reference frame of locally non-rotating observers by the formula $\mathbf{a} = -c^2\nabla \ln\alpha$). Furthermore, the expressions in brackets are 'Lie-type' derivatives (with respect to time) for the set of locally non-rotating observers who move in absolute space and for whom $d\mathbf{x}/dt = \boldsymbol{\beta}$. Thus, these expressions correspond to total derivatives with respect to the times of $\mathbf{E}$ and $\mathbf{B}$, respectively, with the motion of locally non-rotating observers taken into account.

5.5.3.2 Boundary conditions at the event horizon

The Rindler spacetime is a very good approximation in the narrow strip region near the event horizon of a black hole. For this reason, the previous results can be easily generalized to the case of a stationary black hole.

The event horizon is generated by null geodesics which are bicharacteristics of Maxwell's equations. The corresponding boundary conditions at the horizon can be written in a very clear form by introducing a *fictitious* surface electric charge density σ^H which compensates for the flux of the electric field across the surface and a *fictitious* surface electric current $\mathbf{i}^H$ which closes tangent components of the magnetic fields at the horizon. This interpretation is used in the *membrane formalism* [5].

The horizon of a stationary black hole has topology $T \times S^2$, and the surface with the topology S^2 is a two-dimensional surface of infinite gravitational redshift,

$\alpha = 0$. The redshifted gravitational acceleration $\alpha \mathbf{a} \equiv -c^2 \alpha \mathbf{n} \ln \alpha$ remains finite at the horizon:

$$(\alpha \mathbf{a})_H = -\kappa \mathbf{n} \qquad (5.173)$$

where $\mathbf{n}$ is a unit vector pointing orthogonally out of the horizon, and κ is the surface gravity. In the context of calculations near the horizon, it is convenient to introduce a coordinate system (α, λ, ϕ), where λ is a proper distance along the horizon from the north pole toward the equator. In these coordinates, the metric of the absolute three-space near the horizon takes the form

$$ds^2 = (c^2/\kappa)^2 \, d\alpha^2 + d\lambda^2 + \varpi^2 \, d\phi^2 \qquad (5.174)$$

and the unit vectors along the 'toroidal' (ϕ), 'poloidal' (λ), and 'normal' (α) directions are

$$\mathbf{e}_{\hat{\phi}}^\mu \frac{\partial}{\partial x^\mu} = \varpi^{-1} \frac{\partial}{\partial \phi} \qquad \mathbf{e}_{\hat{\lambda}}^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial \lambda} \qquad \mathbf{n}^\mu \frac{\partial}{\partial x^\mu} = \frac{\kappa}{c^2} \frac{\partial}{\partial \alpha}. \qquad (5.175)$$

Macdonald and Thorne [47] formulated the conditions at the horizon as follows:

(i)   Gauss's law: $\mathbf{En} \equiv E_\perp \to 4\pi \sigma^H$;

(ii) charge conservation law: $\alpha \mathbf{jn} \to -\frac{\partial \sigma^H}{\partial t} -^{(2)} \boldsymbol{\nabla} \mathbf{i}^H$;

(iii) Ampere's law: $\alpha \mathbf{B}_\| \to \mathbf{B}^H \equiv (\frac{4\pi}{c}) \mathbf{i}^H \times \mathbf{n}$; and

(iv) Ohm's law: $\alpha \mathbf{E}_\| \to \mathbf{E}^H \equiv R^H \mathbf{i}^H$.

In these relations the symbol $\to$ indicates approach to the black hole horizon along the trajectory of a freely falling observer; $^{(2)}\boldsymbol{\nabla}$ is the two-dimensional divergence at the horizon, and $\mathbf{B}_\|$ and $\mathbf{E}_\|$ are the magnetic and electric field components tangent to the horizon. $R^H \equiv 4\pi/c$ is the effective *surface resistance of the event horizon* ($R^H = 377\Omega$). The lapse function α in the conditions reflects the slowdown in the flow of physical time for locally non-rotating observers in the neighborhood of the black hole.

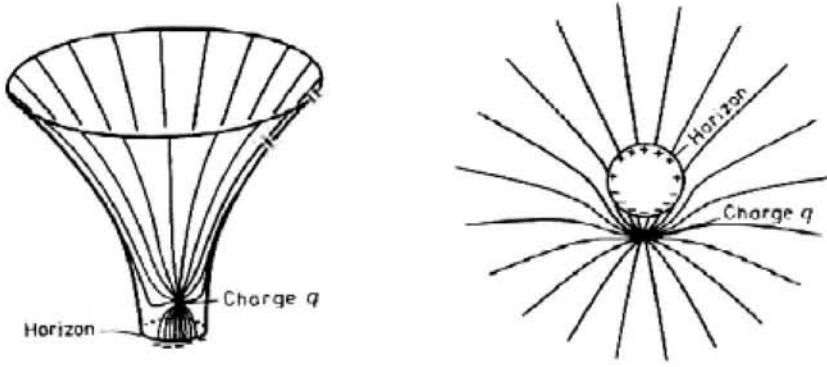### 5.5.4   Electric field of a pointlike charge near a black hole

Let a pointlike charge q be at rest near a Schwarzschild black hole at $r = r_0$, $\theta = 0$. The electric field created by this charge, found by Linet [48] and Léauté and Linet [49], reads:

$$A_\mu(x) = -q\delta_\mu^t \frac{1}{rr_0} \left(M + \frac{\Pi}{R} \right) \qquad (5.176)$$

where

$$R^2 = (r - M)^2 + (r_0 - M)^2 - 2(r - M)(r_0 - M)\lambda - M^2 \sin^2 \theta \qquad (5.177)$$

$$\Pi = (r - M)(r_0 - M) - M^2 \cos\theta. \qquad (5.178)$$

Figure 5.15. The electric-field lines of force of a test charge q at rest in the Schwarzschild metric, in a $\phi = $ constant section: (*left*) lines of force on a curved surface whose geometry coincides with the section $\phi = $ constant of the Schwarzschild metric; (*right*) the same lines projected on a plane ('bird's-eye view'). The distribution of the fictitious surface charge σ^H is shown on the horizon. The charge q is assumed to be positive.

The electric field **E**, defined as $E^i = -\alpha^{-1} g^{ij} A_{t,j}$, is

$$
\begin{aligned}
\mathbf{E} = \frac{q}{r_0 r^2} & \left\{ M \left(1 - \frac{r_0 - M + M \cos\theta}{R} \right) \right. \\
& \left. + \frac{r[(r-M)(r_0-M) - M^2 \cos\theta]}{R^3} (r - M - (r_0 - M)\cos\theta) \right\} \mathbf{e}_{\hat{r}} \\
& + \frac{q(r_0 - 2M)(1 - 2M/r)^{1/2} \sin\theta}{R^3} \mathbf{e}_{\hat{\theta}}
\end{aligned}
\tag{5.179}
$$

where $\mathbf{e}_{\hat{r}}$ and $\mathbf{e}_{\hat{\theta}}$ are unit vectors along the directions of r and θ, respectively.

It is easy to see that at the horizon $\mathbf{E}_\theta \to 0$ so that electric lines of force intersect the horizon at right angles. The total flux of **E** across the horizon is zero (the black hole is uncharged). The pattern of electric lines of force is shown in figure 5.15.

The fictitious charge surface density at the boundary of the black hole is

$$
\sigma^H = \frac{q[M(1 + \cos^2\theta) - 2(r_0 - M)\cos\theta]}{8\pi r_0 [r_0 - M(1 + \cos\theta)]^2}.
\tag{5.180}
$$

Let us bring the charge closer to the horizon ($r_0 \to 2M$). At a distance $r \gg r_0 - 2M$ from the horizon, the lines of force become practically radial and the field strength tends to q/r^2. With the exception of a narrow region close to the horizon, the general picture is almost the same as for a charge placed at the center of the black hole.

5.5.5   Black hole in a magnetic field

5.5.5.1   *Killing vectors and Maxwell fields*

Let us discuss now properties of a black hole immersed in an external magnetic field which is homogenous at infinity. We consider the magnetic field as a test field and neglect its back reaction. This problem allows an elegant solution based on the properties of Killing vector fields. We proceed as follows: a Killing vector in a vacuum spacetime generates a solution of Maxwell's equations [50]. Let us put

$$F_{\mu\nu} = \xi_{\nu;\mu} - \xi_{\mu;\nu} = -2\xi_{\mu;\nu} \tag{5.181}$$

then

$$F^{\mu\nu}{}_{;\nu} = -2\xi^{\mu;\nu}{}_{;\nu} = 0. \tag{5.182}$$

The commutator of two covariant derivatives turns out to be (Ricci identity)

$$\xi_{\mu;\nu;\sigma} - \xi_{\mu;\sigma;\nu} = -\xi^{\lambda} R_{\lambda\mu\nu\sigma}. \tag{5.183}$$

Permuting over the indices μ, ν, and σ, adding the resulting terms, using the Killing equation and the symmetries of the Riemann tensor, one gets

$$\xi_{\mu;\nu;\sigma} = \xi^{\lambda} R_{\lambda\sigma\mu\nu}. \tag{5.184}$$

By contracting the indices ν and σ, we obtain

$$\xi^{\mu;\nu}{}_{;\nu} = R^{\mu}{}_{\lambda}\xi^{\lambda}. \tag{5.185}$$

In a vacuum spacetime we have $R^{\mu}{}_{\lambda} = 0$ and hence $F_{\mu\nu}$ associated with the Killing vector ξ^{μ} satisfies the homogeneous Maxwell equation.

5.5.5.2   *A black hole in a homogeneous magnetic field*

The relation between the Killing vector and an electromagnetic field in Kerr geometry [51] can be used to construct a solution describing a magnetic test field, which is homogeneous at infinity. Let us introduce two fields:

$$F_{(t)\mu\nu} = -2\xi_{(t)\mu;\nu} \qquad F_{(\phi)\mu\nu} = -2\xi_{(\phi)\mu;\nu}. \tag{5.186}$$

At large distances $F_{(t)\mu\nu}$ vanishes, while $F_{(\phi)\mu\nu}$ asymptotically becomes a uniform magnetic field.

It is easy to show that for any two-dimensional surface Σ surrounding a black hole

$$\int_{\Sigma} F_{\mu\nu} \, d\sigma^{\mu\nu} = 0 \tag{5.187}$$

for both fields. Thus the magnetic monopole charge vanishes for both solutions. One also has

$$\int_{\Sigma} {}^*F_{(t)\mu\nu} \, d\sigma^{\mu\nu} = -8\pi M \qquad \int_{\Sigma} {}^*F_{(\phi)\mu\nu} \, d\sigma^{\mu\nu} = 16\pi aM. \tag{5.188}$$

Here $^*F_{(t)\mu\nu} = \varepsilon^{\mu\nu\alpha\beta}F_{\alpha\beta}$. Thus, the axial Killing vector $\xi_{(\phi)}$ generates a stationary, axisymmetric field, which asymptotically approaches a uniform magnetic field and, moreover, has electric charge $4aM$. The timelike Killing vector $\xi_{(t)}$ generates a stationary, axisymmetric field, which vanishes at infinity and has electric charge $-2M$.

Combining these results we conclude that for a neutral black hole the electromagnetic field which asymptotically approaches the homogeneous magnetic field B is given by the vector potential

$$\tilde{A}_\mu = \tfrac{1}{2}B[\xi_{(\phi)\mu} + 2a\xi_{(t)\mu}]. \tag{5.189}$$

The electrostatic *injection energy* per unit charge calculated along the symmetry axis is

$$\epsilon = [\tilde{A}_\mu \xi_{(t)}^\mu]_{r_+}^\infty = -Ba. \tag{5.190}$$

Carter [52] proved that ϵ is constant over the event horizon. Thus a black hole immersed in a rarefied plasma will accrete charge until ϵ vanishes. The resulting black hole charge is

$$Q = 2BaM. \tag{5.191}$$

The vector potential for such a black hole is

$$A_\mu = \tfrac{1}{2}B\xi_{(\phi)\mu}. \tag{5.192}$$

5.5.6   Mechanism of the power generation

5.5.6.1   Potential difference

The potential difference between the (north) pole of the black hole and its equator, as measured by an stationary observer $v^\mu \sim \xi_{(t)}^\mu + \Omega^F \xi_{(\phi)}^\mu$, is

$$\Delta U = U(r_+, \theta = 0) - U(r_+, \theta = \pi/2) \tag{5.193}$$

$$A_\mu(\xi_{(t)}^\mu + \Omega^F \xi_{(\phi)}^\mu) = \tfrac{1}{2}B[g_{t\phi} + \Omega^F g_{\phi\phi}]. \tag{5.194}$$

One has

$$g_{t\phi}(r_+, \theta = 0) = g_{\phi\phi}(r_+, \theta = 0) = 0 \tag{5.195}$$

$$g_{t\phi}(r_+, \theta = \pi/2) = -a \qquad g_{\phi\phi}(r_+, \theta = \pi/2) = r_+^2 + a^2. \tag{5.196}$$

Thus

$$\Delta U = \tfrac{1}{2}B(r_+^2 + a^2)(\Omega^H - \Omega^F) = \tfrac{1}{2}Ba(1 - \Omega^F/\Omega^H). \tag{5.197}$$

For a stationary observer co-rotating with a black hole ΔU vanishes. For $\Omega^F \neq \Omega^H$ there is a non-vanishing electric potential.

One can easily 'predict' this effect by using the analogy of a black hole horizon with a conducting surface with the effective surface resistance $R^H = 4\pi/c = 377\ \Omega$. It is well known that a rotation of a metallic conducting sphere in an external homogeneous magnetic field generates a difference in the potentials between the pole and the equator of the sphere. Such a device is known as a *unipolar inductor*.

5.5.6.2 Black hole magnetosphere and efficiency of the power generating process

Astrophysical black holes are surrounded by plasma. In the most important case for astrophysics, the conductivity of the plasma is so high that the electric field in the reference frame, comoving with the plasma, vanishes, and the magnetic lines of force are 'frozen' into the plasma. In this case, the electric and magnetic fields in an arbitrary reference frame are perpendicular to each other (degenerate fields):

$$\mathbf{E} \cdot \mathbf{B} = 0. \tag{5.198}$$

Note that this condition is only an approximation and generally a small longitudinal electric field is present.   To solve problems concerning the configuration of fields, currents, and charge distributions, it is only necessary that the inequality

$$|\mathbf{E} \cdot \mathbf{B}| \ll |\mathbf{E}^2 - \mathbf{B}^2| \tag{5.199}$$

is satisfied. Small deviations from the exact equation in the neighborhood of a black hole may prove to be important for a number of astrophysical processes.

To simplify the problem, it is usually assumed that the system (a black hole, surrounding plasma, and the electromagnetic field) is stationary and axisymmetric.

Denote by $\tilde{v}^\mu$ a vector of an observer comoving with the plasma. Then one has

$$F_{\mu\nu}\tilde{v}^\nu = 0. \tag{5.200}$$

In the reference frame of this observer, the electric field vanishes. This property is also valid for any frame which is moving with respect to $\tilde{v}^\mu$ with the velocity along the magnetic field. Let us choose a special solution of equation (5.200) which meets the symmetry property

$$v^\mu \sim \xi^\mu_{(t)} + \Omega^F \xi^\mu_{(\phi)} \tag{5.201}$$

where Ω^F is a function of r and θ. Let us stress that the vector $\tilde{v}^\mu$, which has evident physical meaning, must always be timelike, while the vector v^μ can be spacelike. This happens near the horizon if $\Omega^F \approx \Omega^H$.

In the force-free approximation, the rotational energy of the black hole is extracted at a rate of [5]

$$\mathcal{P} = -\frac{d(Mc^2)}{dt} = \int_H \frac{\Omega^F(\Omega^H - \Omega^F)}{4\pi c} \frac{A^H \sin^2\theta}{\Sigma_H} B_\perp^2 \, d\Sigma^H. \tag{5.202}$$

This energy is transferred along magnetic lines of force into a region located far away from the black hole where the force-free condition is violated; energy is pumped into accelerated particles, and so forth.

The power is maximal when $\Omega^F = \Omega^H/2$. Macdonald and Thorne [47] demonstrated that this condition is likely to be implemented in the described model.

In order of magnitude, the power of the 'electric engine' outlined here is

$$\mathcal{P} \approx \left(10^{39}\frac{\text{erg}}{\text{s}}\right)\left(\frac{M}{10^6 M_\odot}\right)^2 \left(\frac{a}{a_{\text{max}}}\right)^2 \left(\frac{B}{10^4\,\text{G}}\right)^2. \qquad (5.203)$$

Here B is the magnetic field strength in the neighborhood of the black hole. Sometimes this electric engine is described in terms of concepts taken from electrical engineering [47,53–56].

5.5.6.3 Black hole as a unipolar inductor

The lines of constant θ are equipotential curves at the horizon since the field $\mathbf{E}^H$ is meridional. Hence, the potential difference between two equipotential lines (marked by 1 and 2) is

$$\Delta U^H = \int_1^2 \mathbf{E}^H\, d\mathbf{l} \approx (10^{17}\,\text{V})\left(\frac{M}{10^6 M_\odot}\right)\left(\frac{B}{10^4\,\text{G}}\right)\left(\frac{a}{a_{\text{max}}}\right) \qquad (5.204)$$

where $d\mathbf{l}$ is the distance element along a meridian of the black hole surface and ΔU^H is the difference between the values of U^H on the equipotentials 1 and 2. The approximate equality is valid if $\Omega^F \approx \Omega^H/2$, maximal Ω^H, and the equipotentials 2 and 1 corresponding to the equatorial and polar regions, respectively.

However, ΔU^H can be written in terms of the surface current $\mathbf{i}^H$ and resistance:

$$\Delta U^H = R^H \mathbf{i}^H \Delta l \qquad (5.205)$$

where Δl is the distance along the meridian between the equipotentials 2 and 1. Substituting the expression for $\mathbf{i}^H$, we obtain

$$\Delta U^H = \frac{I R^H |\Delta l|}{2\pi\varpi_H} = I \Delta Z^H \qquad (5.206)$$

where

$$\Delta Z^H \equiv \frac{R^H |\Delta l|}{2\pi\varpi_H} \qquad (5.207)$$

is the total resistance between the equipotential lines 2 and 1. (If the equipotentials 2 and 1 correspond to the equator and to $\theta \approx \pi/4$, the integration of (5.207) yields $\Delta Z^H \approx 30\,\Omega$.)

These formulas permit the conclusion that in this model the rotating black hole acts as a battery with an e.m.f. of order

$$(10^{17}\,\text{V})\left(\frac{M}{10^6 M_\odot}\right)\left(\frac{B}{10^4\,\text{G}}\right) \qquad (5.208)$$

and an internal resistance of about 30 Ω.

This mechanism (and a number of its variants) has been employed in numerous papers for the explanation of the activity of the nuclei of galaxies and quasars.

References

[1] Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco: Freeman)

[2] Hawking S W and Ellis G F 1973 *The Large-Scale Structure of Spacetime* (Cambridge: Cambridge University Press)

[3] Chandrasekhar S 1983 *The Mathematical Theory of Black Holes* (Oxford: Clarendon)

[4] Wald R M 1984 *General Relativity* (Chicago, IL: University of Chicago Press)

[5] Thorne K S, Price R H and Macdonald D A (eds) 1986 *Black Holes: the Membrane Paradigm* (New Haven, CT: Yale University Press)

[6] Frolov V and Novikov I 1998 *Black Hole Physics: Basic Concepts and New Developments* (Dordrecht: Kluwer)

[7] Boyer R H and Lindquist R W 1967 *J. Math. Phys.* **8** 265

[8] Carter B 1968 *Phys. Rev.* **174** 1559

[9] Carter B 1968 *Commun. Math. Phys.* **10** 280

[10] Carter B 1973a *Black Holes* ed C DeWitt and B S DeWitt (New York: Gordon and Breach)

[11] Carter B 1977 *Phys. Rev.* D **16** 3395

[12] Walker M and Penrose R 1970 *Commun. Math. Phys.* **18** 265

[13] Zakharov A F 1986 *Sov. Phys.–JETP* **64** 1

[14] Zakharov A F 1989 *Sov. Phys.–JETP* **68** 217

[15] Bardeen J M, Press W H and Teukolsky S A 1972 *Astrophys. J.* **178** 347

[16] Penrose R 1969 *Riv. Nuovo Cimento* **1** 252

[17] Dymnikova I G 1986 *Usp. Fiz. Nauk* **148** 393

[18] Young P J 1976 *Phys. Rev.* D **14** 3281

[19] Leaver E W 1986 *Phys. Rev.* D **34** 384

[20] Ching E S C, Leung P T, Suen W M and Young K 1995 *Phys. Rev.* D **52** 2118

[21] Schutz B F and Will C M 1985 *Astrophys. J.* **291** L33

[22] Ferrari V and Mashhoon B 1984 *Phys. Rev. Lett.* **52** 1361

[23] Ferrari V and Mashhoon B 1984 *Phys. Rev.* D **30** 295

[24] Thorne K S 1978 *Theoretical Principles in Astrophysics and Relativity* ed N R Lebovitz, W H Reid and P O Vandervoort (Chicago, IL: University of Chicago Press) p 149

[25] Mashhoon B 1985 *Phys. Rev.* D **31** 290

[26] Price R H 1972 *Phys. Rev.* D **5** 2419

[27] Teukolsky S A 1973 *Astrophys. J.* **185** 635

[28] Cohen J M and Kegeles L S 1974 *Phys. Rev.* D **10** 1070

[29] Chrzanowski P L 1975 *Phys. Rev.* D **11** 2042

[30] Wald R M 1978 *Phys. Rev. Lett.* **41** 203

[31] Press W H and Teukolsky S A 1973 *Astrophys. J.* **185** 649

[32] Teukolsky S A and Press W H 1974 *Astrophys. J.* **193** 443

[33] Chrzanowski P L, Matzner R A, Sandberg V D and Ryan M P Jr 1976 *Phys. Rev.* D **14** 317

[34] Candelas P, Chrzanowski P and Howard K W 1981 *Phys. Rev.* D **24** 297

[35] Gal'tsov D V 1982 *J. Phys. A: Math. Gen.* **15** 3737

[36] Futterman J A H, Handler F A and Matzner R A 1988 *Scattering from Black Holes* (Cambridge: Cambridge University Press)

[37] Fackerell E D and Crossman R G 1977 *J. Math. Phys.* **18** 1849

[38] Detweiler S L 1980 *Phys. Rev.* D **22** 2323

[39] Detweiler S L 1980 *Astrophys. J.* **239** 292

[40] Kojima Y and Nakamura T 1983 *Phys. Lett.* A **99** 37

[41] Kojima Y and Nakamura T 1984 *Prog. Theor. Phys.* **72** 494

[42] Zel'dovich Ya B 1971 *Sov. Phys.–JETP Lett.* **14** 180

[43] Zel'dovich Ya B 1972 *Sov. Phys.–JETP* **35** 1085

[44] Misner C W 1972 *Phys. Rev. Lett.* **28** 994

[45] Starobinsky A A 1973 *Zh. Eksp. Teor. Fiz.* **64** 48 (Engl. Transl. *Sov. Phys.–JETP* **37** 28)

[46] Starobinsky A A and Churilov S M 1973 *Zh. Eksp. Teor. Fiz.* **65** 3

[47] Macdonald D A and Thorne K S 1982 *Mon. Not. R. Astron. Soc.* **198** 345

[48] Linet B 1976 *J. Phys. A: Math. Gen.* **9** 1081

[49] Léauté B and Linet B 1976 *Phys. Lett.* A **58** 5

[50] Papapertrou A 1966 *Ann. Inst. Henri Poincaré* A **4** 83

[51] Wald R M 1974 *Phys. Rev.* D **10** 1680

[52] Carter B 1973 *Black Holes* ed C DeWitt and B S DeWitt (New York: Gordon and Breach)
Carter B 1973 *Commun. Math. Phys.* **30** 261

[53] Blandford R D 1979 *Active Galactic Nuclei* ed C Hazard and S Mitton (Cambridge: Cambridge University Press) p 241

[54] Znajek R L 1978 *Mon. Not. R. Astron. Soc.* **185** 833

[55] Damour T 1978 *Phys. Rev.* D **18** 3598

[56] Thorne K S and Blandford R D 1982 *Extragalactic Radio Sources* ed D Heeschen and C Wade p 255

Chapter 6

Close encounters of black holes

Domenico Giulini
University of Freiburg, Germany

Processes involving close encounters of black holes, like their mutual scattering or merging, can be analytically approached by Einstein's equations without additional matter. Some characteristic features of these equations are reviewed and set in relation to Newtonian concepts. The initial-value problem is explained and techniques are presented that allow the explicit construction of multi-black-hole initial data. Some physical properties of these data are discussed and, finally, a perspective on recent developments is given.

6.1 Introduction and motivation

In my chapter I will try to explain how scattering and merging processes between black holes can be described analytically in general relativity (GR). This is a vast subject and I will focus attention on the basic issues, rather than trying to explain the analytical details of approximation schemes etc. I will also not discuss numerical aspects, which are beyond my competence, and which would anyway require a separate chapter. I will address the following main topics:

(1) a first step beyond Newtonian gravity,
(2) constrained evolutionary structure of Einstein's equations,
(3) the 3 + 1 split and the Cauchy initial-value problem,
(4) black hole data and
(5) problems and recent developments

with emphasis on the fourth entry. However, I will also spend some time in explaining some of the specialties of GR, like the absence of a point-particle concept and the non-trivial linkage between the field equations and the equations of motion for matter. These points should definitely be appreciated before one

178

goes on to discuss black holes, which are solutions to the *vacuum* Einstein equations representing *extended* objects.

The following points seem to me the main motivations for studying the problem of black hole collision:

- Coalescing black holes are regarded as promising sources for the detection of gravitational waves by earth-based instruments.
- Close encounters of black holes provide physically relevant situations for the investigation of the strong-field regime of general relativity.
- The dynamics of simple black hole configurations is regarded as an ideal testbed for numerical relativity.

My conventions are as follows: spacetime is a manifold M with Lorentzian metric g of signature $(-, +, +, +)$. Greek indices range from 0 to 3, Latin indices from 1 to 3 unless stated otherwise. The covariant derivative is denoted ∇_μ, ordinary partial derivatives by ∂_μ or sometimes simply by a lower-case μ. The relation $:=$ $(=:)$ defines the left- (right-) hand side. The gravitational constant in GR is $\kappa = 8\pi G/c^2$, where G is Newton's constant and c the velocity of light. A symbol like $O(\epsilon^n)$ stands collectively for terms falling off at least as fast as ϵ^n.

6.2 A first step beyond Newtonian gravity

It can hardly be overstressed how useful the concept of a point mass is in Newtonian mechanics and gravity. It allows us to probe the gravitational field *pointwise* and to reduce the dynamical problem to the mathematical problem of finding solutions to a system of *finitely* many *ordinary* differential equations. To be sure, just postulating the existence of point masses is not sufficient. To be consistent with the known laws of physics, one must eventually understand the point mass as an idealization of a highly localized mass distribution which obeys known field-theoretic laws, such that in the situations at hand most of the field degrees of freedom effectively decouple from the dynamical laws for those collective degrees of freedom in which one is interested, e.g. the centre of mass. In Newtonian gravity this usually requires clever approximation schemes but is not considered to be a problem of fundamental nature. Although this is true for the specific linear theory of Newtonian gravity, this need not be so for comparably simple generalizations, as will become clear later.

In GR the situation is markedly different. A concentration of more than one Schwarzschild mass in a region of radius less than the Schwarzschild radius will lead to a black hole whose behaviour away from the stationary state cannot usually be well described by finitely many degrees of freedom. It shakes and vibrates, thereby radiating off energy and angular momentum in the form of gravitational radiation. Moreover, it is an extended object and cannot be unambiguously ascribed an (absolute or relative) position or individual mass. Hence the problem of motion, and therefore the problem of scattering of black

holes, cannot be expected to merely consist of *corrections* to Newtonian scattering problems. Rather, the whole kinematic and dynamical setup will be different where many of the established concepts of Newtonian physics need to be replaced or at least adapted, very often in a somewhat ambiguous way. Among these are mass, distance, and kinetic energy. For example, one may try to solve the following straightforward sounding problem in GR, whose solution one might think has been given long ago. Consider two unspinning black holes, momentarily at rest, with equal individual mass m, mutual distance ℓ, and no initial gravitational radiation around. What is the amount of energy released via gravitational radiation during the dynamical infall? In such a situation we can usually make sense of the notions of 'spin' (hence unspinning) and 'mass'; but ambiguities generally exist in defining 'distance' and, most important of all, 'initial gravitational radiation'. Such difficulties persist over and above the ubiquitous analytical and/or numerical problems which are currently under attack by many research groups.

To those who are not so familiar with GR and like to see Newtonian analogies, I wish to mention that there is a way to *consistently* model some of the nonlinear features of Einstein's equations in a Newtonian context, which shares the property that it does not allow for point masses. I will briefly describe this model since it does not seem to be widely known.

First recall the field equation in Newtonian gravity, which allows us to determine the gravitational potential ϕ (whose negative gradient, $-\vec{\nabla}\phi$, is the gravitational field) from the mass density ρ (the 'source' of the gravitational field):

$$\Delta\phi = 4\pi G\rho. \tag{6.1}$$

Now suppose one imposes the following principle for a modification of (6.1): all energies, including the self-energy of the gravitational field, act as source for the gravitational field. In order to convert an energy density ε into a mass density ρ, we adopt the relation $\varepsilon = \rho c^2$ from special relativity (the equation we will arrive at can easily be made Lorentz invariant by adding appropriate time derivatives). The question then is whether one can modify the source term of (6.1) such that $\rho \to \rho + \rho_{grav}$ with $\rho_{grav} := \varepsilon_{grav}/c^2$, where ε_{grav} is the energy density of the gravitational field *as predicted by this very same equation* (condition of self-consistency). It turns out that there is indeed a unique such modification, which reads:

$$\Delta\phi = \frac{4\pi G}{c^2}\phi\left[\rho + \frac{c^2}{8\pi G}(\vec{\nabla}\phi/\phi)^2\right]. \tag{6.2}$$

It is shown in the appendix at the end of the chapter that this equation indeed satisfies the 'energy principle' as just stated. (For more information and a proof of uniqueness, see [21].) The gravitational potential is now required to be always positive, tending to the value c^2 at spatial infinity (rather than zero as for (6.1)). The second term on the right-hand side of (6.2) corresponds to the energy density of the gravitational field. Unlike the energy density following from (6.1) (which

is $-\frac{1}{8\pi G}|\nabla\phi|^2$) it is now *positive* definite. This does not contradict the attractivity of gravity for the following reason: the rest-energy density of a piece of matter is in this theory not given by ρc^2, but by $\rho\phi$, that is, it depends on the value of the gravitational potential at the location of the matter. The same piece of matter located at a lower gravitational potential has less energy than at higher potential values. In GR this is called the universal redshift effect. Here, as in GR, the active gravitational mass also suffers from this redshift, as is immediate from the first term on the right-hand side of (6.2), where ρ does not enter alone, as in (6.1), but is multiplied with the gravitational potential ϕ. With respect to these features our modification (6.2) of Newtonian gravity mimics GR quite well.

We mention in passing that (6.2) can be 'linearized' by introducing the dimensionless field $\psi := \sqrt{\phi}/c$, in terms of which (6.2) reads:

$$\Delta\psi = \frac{2\pi G}{c^2}\rho\psi. \tag{6.3}$$

The boundary conditions are now $\psi(r \to \infty) \to 1$. Hence only those linear combinations of solutions whose coefficients add up to one are again solutions. For $\rho \geq 0$ it also follows that solutions to (6.3) can never assume negative values, since otherwise the function ψ must have a negative minimum (because of the positive boundary values) and therefore non-negative second derivatives are there. But then (6.3) cannot be satisfied at the minimum, hence ψ must be non-negative everywhere. This implies that solutions of (6.2) are also non-negative. To be sure, for mathematical purposes (6.3) is easier to use than (6.2), but note that ϕ and not ψ is the physical gravitational potential.

We now show how these nonlinear features render impossible the notion of a point mass, and even induce a certain black hole behaviour on their solutions. Let us be interested in static, spherically symmetric solutions to (6.2) with source ρ, which is zero for $r > R$ and constant for $r < R$. We need to distinguish two notions of mass. One mass just counts the amount of 'stuff' located within $r < R$. You may call it the 'bare mass' or 'baryonic mass', since for ordinary matter it is proportional to the baryon number. We denote it by M_B. It is simply given by

$$M_\text{B} := \int_\text{space} d^3x\,\rho. \tag{6.4}$$

The other mass is the 'gravitational mass', which is measured by the amount of flux of the gravitational field to 'infinity', that is, through the surface of a sphere whose radius tends to infinity. We call this mass M_G. It is given by

$$M_\text{G} := \frac{1}{4\pi G}\lim_{r\to\infty}\int_{S^2(r)}(-\vec{\nabla}\phi\cdot\vec{n})\,d\sigma \tag{6.5}$$

where $r = |\vec{x}|$, $\vec{x}/r = \vec{n}$, $S^2(r)$ is the two-sphere of radius r and $d\sigma$ is its surface element. M_G should be identified with the total inertial mass of the system, in full analogy to the ADM mass in GR (see equation (6.35)). Hence

$M_G c^2$ is the total energy of the system, with gravitational binding energy also taken into account. The masses M_B and M_G can dimensionally be turned into radii by writing $R_B := GM_B/c^2$ and $R_G := GM_G/c^2$, and further turned into dimensionless quantities via rescaling with R, the radius of our homogeneous star. We write $x := R_B/R$ and $y := R_G/R$.

For each pair of values for the two parameters M_B and R there is a unique homogeneous-star solution to (6.2), whose simple analytical form need not interest us here (see [21]). Using it we can calculate M_G, whose dependence on the parameters is best expressed in terms of the dimensionless quantities x and y:

$$y = f(x) = 2 \left(1 - \frac{\tanh(\sqrt{3x/2})}{\sqrt{3x/2}} \right). \tag{6.6}$$

The function f maps the interval $[0, \infty]$ monotonically to $[0, 2]$. This implies the following inequality

$$M_G < 2Rc^2/G \tag{6.7}$$

which says that the gravitational mass of the star is bounded by a purely geometric quantity. It corresponds to the statement in GR that the star's radius must be bigger than its Schwarzschild radius, which in isotropic coordinates is indeed given by $R_S = GM_G/2c^2$. It can be proven [21] that the bound (6.7) still exists for non-homogeneous spherically symmetric stars, so that the somewhat unphysical homogeneity assumption can be lifted. The physical reason for this inequality is the 'redshift', i.e. the fact that the same bare mass at lower gravitational potential produces less gravitational mass. Hence adding more and more bare mass into the same volume pushes the potential closer and closer to zero (recall that ϕ is always positive) so that the added mass becomes less and less effective in generating gravitational fields. The inequality then expresses the mathematical fact that this 'redshifting' is sufficiently effective so as to give finite upper bounds to the gravitational mass, even for unbounded amounts of bare mass.

The energy balance can also be nicely exhibited. Integrating the matter energy density $\phi\rho$ and the energy density of the gravitational field, $\frac{c^4}{8\pi G}(\vec{\nabla}\phi/\phi)^2$, we obtain

$$E_{\text{matter}} = M_B c^2 \left(1 - \frac{6R_B}{5R} + O(R_B^2/R^2) \right) \tag{6.8}$$

$$E_{\text{field}} = M_B c^2 \left(\frac{3R_B}{5R} + O(R_B^2/R^2) \right) \tag{6.9}$$

$$E_{\text{total}} = M_B c^2 \left(1 - \frac{3R_B}{5R} + O(R_B^2/R^2) \right) = M_G c^2. \tag{6.10}$$

Note that the term $-3M_B c^2 R_B/5R$ in (6.10) is just the Newtonian binding energy. At this point it is instructive to verify the remarks we made earlier about the positivity of the gravitational energy. Shrinking a mass distribution *enhances* the field energy, but diminishes the matter energy twice as fast, so that the overall

energy is also diminished, as it must be due to the attractivity of gravity. But here this is achieved with all involved energies being positive, unlike in Newtonian gravity. Note that the total energy, M_G, cannot become negative (since ϕ cannot become negative, as has already been shown). Hence one also cannot extract an infinite amount of energy by unlimited compression, as is possible in Newtonian gravity. This is the analogue in our model theory to the positive mass theorem in GR.

We conclude by making the point announced earlier, namely that the inequality (6.7) shows that point objects of finite gravitational mass do not exist in the theory based upon (6.2); mass implies extension! Taken together with the lesson from special relativity, that extended rigid bodies also do not exist (since the speed of elastic waves is less than c), we arrive at the conclusion that the dynamical problem of gravitating bodies and their interaction is fundamentally field theoretic (rather than point mechanical) in nature. Its proper realization is GR to which we now turn.

6.3   Constrained evolutionary structure of Einstein's equations

In GR the basic field is the spacetime metric $g_{\mu\nu}$, which comprises the gravitational and inertial properties of spacetime. It defines what inertial motion is, namely a geodesic

$$\ddot{x}^\lambda + \Gamma^\lambda_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0 \tag{6.11}$$

with respect to the Levi-Civita connection

$$\Gamma^\lambda_{\mu\nu} := \tfrac{1}{2}g^{\lambda\sigma}(-g_{\mu\nu,\sigma} + g_{\sigma\mu,\nu} + g_{\nu\sigma,\mu}). \tag{6.12}$$

(Since inertial motion is 'force free' by definition, you may rightly ask whether it is correct to call gravity a 'force'.) The gravitational field $g_{\mu\nu}$ is linked to the matter content of spacetime, represented in the form of the energy–momentum tensor $T_{\mu\nu}$, by Einstein's equations

$$G_{\mu\nu} := R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = \kappa T_{\mu\nu}. \tag{6.13}$$

Due to the gauge invariance with respect to general differentiable point transformations (i.e. diffeomorphisms) of spacetime, one has the *identities* (as a consequence of Noether's second theorem)

$$\nabla_\mu G^{\mu\nu} \equiv 0. \tag{6.14}$$

Being 'identities' they hold for any $G^{\mu\nu}$, independent of any field equation. With respect to some coordinate system $x^\mu = (x^0, \ldots x^3)$ we can expand (6.14) in terms of ordinary derivatives. Preferring the coordinate x^0, this reads:

$$\partial_0 G^{0\nu} = -\partial_k G^{k\nu} - \Gamma^\mu_{\mu\lambda}G^{\lambda\nu} - \Gamma^\nu_{\mu\lambda}G^{\mu\lambda}. \tag{6.15}$$

Since $G^{\mu\nu}$ contains no higher derivatives of $g_{\mu\nu}$ than the second, the right-hand side of this equation also contains only second x^0 derivatives. Hence (6.15) implies that the four components $G^{0\nu}$ only involve first x^0-derivatives. Now choose x^0 the as time coordinate. The four $(0, \nu)$-components of (6.13) then do not involve second time derivatives, unlike the space–space components (i, j). Hence the time–time and time–space components are *constraints*, that is, equations that constrain the allowed choices of initial data, rather than evolving them.

This is not an unfamiliar situation as it similarly occurs for Maxwell's equations in electromagnetism (EM). Let us recall this analogy. We consider the four-dimensional form of Maxwell's equations in terms of the vector potential A_μ, whose antisymmetric derivative is the field tensor $F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu$, comprising the electric $(E_i = F_{0i})$ and magnetic $(B_i = -F_{jk}, ijk$ cyclic) fields. Maxwell's equations are

$$E^\nu := \partial_\mu F^{\mu\nu} - \frac{4\pi}{c} j^\nu = 0. \tag{6.16}$$

Due to its antisymmetry, the field tensor obviously obeys the identity

$$\partial_\mu \partial_\nu F^{\mu\nu} \equiv 0 \tag{6.17}$$

which here is the analogue of (6.14), an identity involving third derivatives in the field variables. Using (6.17) in the divergence of (6.16) yields

$$\partial_\nu E^\nu = -\frac{4\pi}{c} \partial_\nu j^\nu \tag{6.18}$$

which shows that Maxwell's equations imply charge conservation as the integrability condition. Let us interpret the rôle of charge conservation in the initial-value problem. Decomposing (6.17) into space and time derivatives gives

$$\partial_0 \partial_\nu F^{0\nu} = -\partial_k \partial_\nu F^{k\nu}. \tag{6.19}$$

Again the right-hand side involves only second time derivatives implying that the zero-component of (6.16) involves no second time derivatives. Hence the time component of (6.16) is merely a constraint on the initial data; clearly it is just Gauss's law $\vec{\nabla} \cdot \vec{E} - 4\pi\rho = 0$. Its change under time evolution according to Maxwell's equations is

$$\partial_0 E^0 = \partial_\nu E^\nu - \partial_k E^k$$
$$= -\frac{4\pi}{c} \partial_\nu j^\nu - \partial_k E^k \tag{6.20}$$

where we have used the identity (6.18) in the second step. Suppose now that on the initial surface of constant x^0 we put an electromagnetic field which satisfies the constraint, $E^0 = 0$, and which we evolve according to $E^k = 0$ (implying

$\partial_k E^k = 0$ on that initial surface). Then (6.20) shows that charge conservation is a necessary and sufficient condition for the evolution to preserve the constraint.

Let us return to GR now, where the overall situation is entirely analogous. Now we have four constraints

$$E^{0\nu} := G^{0\nu} - \kappa T^{0\nu} = 0 \tag{6.21}$$

and six evolution equations, which we write

$$E^{ij} := G^{ij} - \kappa T^{ij} = 0. \tag{6.22}$$

The identity (6.14) now implies

$$\nabla_\mu E^{\mu\nu} = -\kappa \nabla_\mu T^{\mu\nu} \tag{6.23}$$

which parallels (6.18). Here, too, the time derivative of the constraints is easily calculated,

$$
\begin{aligned}
\partial_0 E^{0\nu} &= \nabla_\mu E^{\mu\nu} - \partial_k E^{k\nu} - \Gamma^0_{0\lambda} E^{\lambda\nu} - \Gamma^\nu_{0\lambda} E^{0\lambda} \\
&= -\kappa \nabla_\mu T^{\mu\nu} - \partial_k E^{k\nu} - \Gamma^0_{0\lambda} E^{\lambda\nu} - \Gamma^\nu_{0\lambda} E^{0\lambda}
\end{aligned}
\tag{6.24}
$$

by using (6.23) in the last step. Now consider again the evolution of initial data from a surface of constant x^0. If they initially satisfy the constraints and are evolved via $E^{ij} = 0$ (hence all spatial derivatives of $E^{\mu\nu}$ vanish initially) they continue to satisfy the constraints if and only if the energy–momentum tensor of the matter satisfies

$$\nabla_\mu T^{\mu\nu} = 0. \tag{6.25}$$

Hence we see that the 'covariant conservation' of energy–momentum, expressed by (6.25), plays the same rôle in GR as charge conservation plays in EM. This means that you cannot just prescribe the motion of matter and then use Einstein's equations to calculate the gravitational field produced by that source. You have to move the matter in such a way that it satisfies (6.25). But note that at this point there is a crucial mathematical difference to charge conservation in EM: charge conservation is a condition on the source *only*, it does not involve the electromagnetic field. This means that you know *a priori* what to do in order not to violate charge conservation. However, (6.25) involves the source *and* the gravitational field. The latter enters through the covariant derivatives which involve the metric $g_{\mu\nu}$ through the connection coefficients (6.12). Hence here (6.25) cannot be solved *a priori* by suitably restricting the motion of the source. Rather we have a consistency condition which *mutually* links the problem of motion for the sources and the problem of field determination. It is this difference which makes the problem of motion in GR exceedingly difficult. (A brief and lucid presentation of this problem, drawing attention to its relevance in calculating the generation of gravitational radiation by self-gravitating systems, was given in [15]. A broader summary, including modern developments, is [14].)

For example, for pressureless dust represented by $T^{\mu\nu} = \rho c^2 \, U^\mu U^\nu$, where ρ is the local rest-mass density and U^μ is the vector field of four-velocities of the continuously dispersed individual dust grains, (6.25) is equivalent to the two equations:

$$\nabla_\mu(\rho U^\mu) = 0 \tag{6.26}$$

$$U^\nu \nabla_\nu U^\mu = 0. \tag{6.27}$$

The first states the conservation of rest-mass. The second is equivalent to the statement that the vector field U^μ is geodesic, which means that its integral lines (the world lines of the dust grains) are geodesic curves (6.11) *with respect to the metric* $g_{\mu\nu}$. Hence we see that in this case the motion of matter is fully determined by (6.25), i.e. by Einstein's equations, which imply (6.25) as integrability condition. This clearly demonstrates how the problem of motion is inseparably linked with the problem of field determination and that these problems can only be solved simultaneously. The methods used today use clever approximation schemes. For example, one can make use of the fact that there is a difference of one power in κ between the field equations and their integrability condition. Hence, in an approximation in κ, it is consistent for the nth-order approximation of the field equations to have the integrability conditions (equations of motions) satisfied to $n - 1$st order.

Clearly the problem just discussed does not arise for the matter-free Einstein equations for which $T_{\mu\nu} \equiv 0$. Now recall that black holes are described by matter-free equations, too. Hence, the mathematical problem just described does not occur in the discussion of their dynamics. In *this* aspect the discussion of black hole scattering is considerably easier than, e.g., that of neutron stars.

6.4   The 3 + 1 split and the Cauchy initial-value problem

We saw that the ten Einstein equations decompose into two sets of four and six equations respectively, four constraints which the initial data have to satisfy, and six equations driving the evolution. As a consequence, there will be four dynamically undetermined components among the ten components of the gravitational field $g_{\mu\nu}$. The task is to parametrize the $g_{\mu\nu}$ in such a way that four dynamically undetermined functions can be cleanly separated from the other six. One way to achieve this is via the splitting of spacetime into space and time (see [22] for a more detailed discussion). The four dynamically undetermined quantities will be the famous *lapse* (one function α) and *shift* (three functions β^i). The dynamically determined quantity is the Riemannian metric h_{ij} on the spatial three-manifolds of constant time. These together parametrize $g_{\mu\nu}$ as follows:

$$ds^2 = -\alpha^2 (dx^0)^2 + h_{ik}(dx^i + \beta^i \, dx^0)(dx^k + \beta^k \, dx^0). \tag{6.28}$$

The physical interpretation of α and β^i is: think of spacetime as the history of space. Each 'moment' of time, $x^0 = t$, corresponds to an entire three-dimensional

slice Σ_t. Obviously there is plenty of freedom in how to 'waft' space through spacetime. This freedom corresponds precisely to the freedom to choose the 1+3 functions α and β^i. For one thing, you may freely specify how far for each parameter step dt you push space in a perpendicular direction forward in time. This is controlled by α, which is just the ratio ds/dt of the *proper* perpendicular distance between the hypersurfaces Σ_t and Σ_{t+dt}. This speed may be chosen in a space- and time-dependent fashion, which makes α a function on spacetime. Second, let a point be given with coordinates x^i on Σ_t. Going from x^i in a perpendicular direction you meet Σ_{t+dt} in a point with coordinates $x^i + dx^i$, where dx^i can be chosen at will. This freedom of moving the coordinate system around while evolving is captured by β^i; one writes $dx^i = \beta^i dt$. Clearly this moving around of the spatial coordinates can also be made in a space- and time-dependent fashion, so that the β^i are functions of spacetime, too.

Let n^μ be the vector field in a spacetime which is normal to the spatial sections of constant time. It is given by $n = \frac{1}{\alpha}(\partial/\partial x^0 - \beta^i \partial/\partial x^i)$, as one may readily verify by using (6.28) (you have to check that n is normalized and satisfies $g(n, \partial/\partial x^i) = 0$). We define the *extrinsic curvature*, K_{ij}, to be one-half the Lie derivative of the spatial metric in the direction of the normal:

$$K_{ij} := \frac{1}{2} L_n h_{ij} = \frac{1}{2\alpha} \left(\frac{\partial h_{ij}}{\partial x^0} - 2D_{(i}\beta_{j)} \right) \tag{6.29}$$

where D is the spatial covariant derivative with respect to the metric h_{ij}. As usual, a round bracket around indices denotes their symmetrization. Note that, by definition, K_{ij} is symmetric. Finally we denote the Ricci scalar of h_{ij} by $R^{(3)}$.

We can now write down the four constraints of the vacuum Einstein equations in terms of these variables:

$$0 = G(n, n) = \frac{1}{2}(R^{(3)} + K^{ij} K_{ij} - (K_i^i)^2) \tag{6.30}$$

$$0 = G(n, \partial/\partial x^j) = D_i(K_j^i - \delta_j^i K_k^k). \tag{6.31}$$

Equations (6.30) and (6.31) are referred to as the *Hamiltonian constraint* and *momentum constraint*, respectively. The six evolution equations of second order in the time derivative can now be written as 12 equations of first order. Six of them are just (6.29), read as the equation that relates the time derivative $\partial h_{ij}/\partial x^0$ to the 'canonical data' (h_{ij}, K_{ij}). The other six equations, whose explicit form needs not concern us here (see e.g. [22]), express the time derivative of K_{ij} in terms of the canonical data. Both sets of evolution equations contain, on their right-hand sides, the lapse and shift functions, whose evolution is not determined but must be specified by hand. This specification is a choice of gauge, without which one cannot determine the evolution of the physical variables (h_{ij}, K_{ij}).

The initial-data problem takes now the following form:

(1) Choose a topological three-manifold Σ.
(2) Find on Σ a Riemannian metric h_{ij} and a symmetric tensor field K_{ij} which satisfy the constraints (6.30) and (6.31).

(3) Choose a lapse function α and a shift vector field β^i, both as functions of space and time, possibly according to some convenient prescription (e.g. singularity avoiding gauges, like maximal slicing).

(4) Evolve initial data with these choices of α and β^i according to the 12 equations of first order. By consistency of Einstein's equations, the constraints will be preserved during this evolution, independent of the choices for α and β^i.

The backbone of this setup is a mathematical theorem, which states that for any set of initial data, taken from a suitable function space, there is, up to a diffeomorphism, a unique maximal Einstein spacetime developing from these data [7].

6.5 Black hole data

6.5.1 Horizons

By black hole data we understand vacuum data which contain *apparent horizons*. The informal definition of an apparent horizon is that it is the boundary of a trapped region, which means that its orthogonal outgoing null rays must have zero divergence. (Inside the trapped region they converge for any two-surface, by the definition of a 'trapped region'.) The Penrose–Hawking singularity theorems state that the existence of an apparent horizon implies that the evolving spacetime will be singular (assuming the strong energy condition). Given also the condition that singularities cannot be seen by observers far off, a condition usually called *cosmic censorship*, one infers the existence of an *event horizon* and hence a black hole. One can then show that the intersection of the event horizon with the spatial hypersurface lies on or outside the apparent horizon (for stationary spacetimes they coincide). The reason why one does not deal with event horizons directly is that one cannot tell whether one exists by just looking at initial data. In principle one would have to evolve them to the infinite future, which is beyond our abilities in general. In contrast, apparent horizons can be recognized once the data on an initial slice are given. The formal definition of an apparent horizon is the following: given initial data (Σ, h_{ij}, K_{ij}) and an embedded two-surface $\sigma \subset \Sigma$ with outward pointing normal v^i, σ is an apparent horizon if and only if the following relation between K_{ij}, the extrinsic curvature of Σ in spacetime, and k_{ij}, the extrinsic curvature of σ in Σ, is satisfied,

$$q^{ij}k_{ij} = -q^{ij}K_{ij} \qquad (6.32)$$

where $q_{ij} := h_{ij} - v_i v_j$ is just the induced Riemannian metric on σ, so that (6.32) simply says that the restriction of K_{ij} to the tangent space of σ has opposite trace to k_{ij}. (The minus sign on the right-hand side of (6.32) signifies a *future* apparent horizon corresponding to a *black* hole which has a *future* event horizon. A plus sign would signify a past apparent horizon corresponding to a '*white* hole' with

past event horizon.) This means that once we have the data (Σ, h_{ij}, K_{ij}) we can in principle find all two-surfaces $\sigma \subset \Sigma$ for which (6.32) holds and therefore find all apparent horizons.

6.5.2 Poincaré charges

By Poincaré charges we shall understand quantities like mass, linear momentum, and angular momentum. In GR they are associated with an asymptotic Poincaré symmetry (see [4]), provided that the data (Σ, h_{ij}, K_{ij}) are *asymptotically flat* in a suitable sense, which we will now explain. Topologically asymptotic flatness means that the non-trivial 'topological features' of Σ should all reside in a bounded region and not 'pile up' at infinity. More formally this is expressed by saying that there is a bounded region $B \subset \Sigma$ such that $\Sigma - B$ (the complement of B) consists of a finite number of disjoint pieces, each of which looks topologically like the complement of a ball in $\mathbb{R}^3$. These asymptotic pieces are also called the *ends* of the manifold Σ. Next comes the geometric restriction imposed by the condition of asymptotic flatness. It states that for each end there is an asymptotically Euclidean coordinate system $\{x^1, x^2, x^3\}$ in which the fields (h_{ij}, K_{ij}) have the following fall-off for $r \to \infty$ $(r = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2}$, $n^k = x^k/r)$:

$$h_{ij}(x^k) = \delta_{ij} + \frac{s_{ij}(n^k)}{r} + O(r^{-1-\epsilon}) \tag{6.33}$$

$$K_{ij}(x^k) = \frac{t_{ij}(n^k)}{r^2} + O(r^{-2-\epsilon}). \tag{6.34}$$

Moreover, in order to have convergent expressions for physically relevant quantities, like, e.g., angular momentum (see later), the field s_{ij} must be an *even* function of its argument, i.e. $s_{ij}(-n^k) = s_{ij}(n^k)$, and t_{ij} must be an *odd* function, i.e. $t_{ij}(-n^k) = -t_{ij}(n^k)$.

Under these conditions each end can be assigned mass, momentum, and angular momentum, which are conserved during time evolution. They may be computed by integrals over two-spheres in the limit that the spheres are pushed to larger and larger radii into the asymptotically flat region of that end. These so-called ADM integrals (first considered by Arnowitt, Deser, and Misner in [1]) are given by the following expressions, which we give in 'geometric' units (meaning that in order to get them in standard units one has to multiply the mass expression given below by $1/\kappa$ and the linear and angular momentum by c/κ):

$$M = \lim_{r \to \infty} \int_{S^2(r)} \delta^{ij}(\partial_i h_{jk} - \partial_k h_{ij})n^k \, d\sigma \tag{6.35}$$

$$P^i = \lim_{r \to \infty} \int_{S^2(r)} (K_k^i - \delta_k^i K_j^j)n^k \, d\sigma \tag{6.36}$$

$$S^i = \lim_{r \to \infty} \int_{S^2(r)} \varepsilon_{ijl} x^j (K_k^l - \delta_k^l K_n^n)n^k \, d\sigma. \tag{6.37}$$

6.5.3   Maximal and time-symmetric data

The constraints (6.30), (6.31) are too complicated to be solved in general. Further conditions are usually imposed to reduce the complexity of the problem: data (h_{ij}, K_{ij}) are called *maximal* if $K_i^i = h^{ij} K_{ij} = 0$. The name derives from the fact that $K_i^i = 0$ is the necessary and sufficient condition for a hypersurface to have a stationary volume to first order with respect to deformations in the ambient spacetime. Even though stationarity does not generally imply extremality, one calls such hypersurfaces maximal. Note also that since spacetime is a Lorentzian manifold, extremal spacelike hypersurfaces will be of maximal rather than minimal volume. In contrast, in Riemannian manifolds one would speak of minimal surfaces.

A much stronger condition is to impose $K_{ij} = 0$ which, as seen from (6.36) and (6.37), implies that all momenta and angular momenta vanish. Only the mass is now allowed to be non-zero. Such data are called *time symmetric* since for them h_{ij} is momentarily static as seen from (6.29). This implies that the evolution of such data into the future and into the past will coincide so that the developed spacetime will have a time-reversal symmetry which pointwise fixes the initial surface where $K_{ij} = 0$. This surface is therefore also called the *moment of time symmetry*. Time-symmetric data can still represent configurations of any number of black holes without angular momenta which are momentarily at rest. Note also that for time-symmetric data, condition (6.32) for an apparent horizon is equivalent to the tracelessness of the extrinsic curvature of σ. Hence for time-symmetric data apparent horizons are minimal surfaces.

We add one more general comment concerning submanifolds. A vanishing extrinsic curvature is equivalent to the property that each geodesic of the ambient space, which starts on, and tangent to the submanifold, will always run entirely inside the submanifold. Therefore, submanifolds with vanishing extrinsic curvature are called *totally geodesic*. Now, if the ambient space allows for an isometry (symmetry of the metric), whose fixed-point set is the submanifold in question, as for the time-reversal transformation just discussed, the submanifold must necessarily be totally geodesic. To see this, consider a geodesic of the ambient space which starts on, and tangent to the submanifold. Assume that this geodesic eventually leaves the submanifold. Then its image under the isometry would again be a geodesic (since isometries always map geodesics to geodesics) which is different from the one from which we started. But this is impossible since they share the same initial conditions which are known to determine the geodesic uniquely. Hence the geodesic cannot leave the submanifold, which proves the claim. We will later have more opportunities to identify totally geodesic submanifolds—namely apparent horizons—by their property of being fixed-point sets of isometries.

6.5.4 Solution strategy for maximal data

Possibly the most popular approach to solving the constraints is the *conformal technique* due to York *et al* (see [35] for a review). The basic idea is to regard the Hamiltonian constraint (6.30) as an equation for the conformal factor of the metric h_{ij} and *freely* specify the complementary information, called the conformal equivalence class of h_{ij}. More concretely, this works as follows:

(1) Choose unphysical ('hatted') quantities $(\hat{h}_{ij}, \hat{K}_{ij})$, where $\hat{h}_{ij}$ is a Riemannian metric on Σ and $\hat{K}_{ij}$ is symmetric, trace and divergence free:

$$\hat{h}^{ij} \hat{K}_{ij} = 0 \qquad \hat{D}^i \hat{K}_{ij} = 0 \qquad (6.38)$$

where $\hat{D}$ is the covariant derivative with respect to $\hat{h}_{ij}$.

(2) Solve the (quasilinear elliptic) equation for a positive, real valued function Φ with boundary condition $\Phi\ r \to \infty) \to 1$, where $\hat{\Delta} = \hat{h}^{ij} \hat{D}_i \hat{D}_j$:

$$\hat{\Delta}\Phi + \tfrac{1}{8}\hat{K}^{ij} \hat{K}_{ij} \Phi^{-7} = 0. \qquad (6.39)$$

(3) Using the solution of (6.39), define physical ('unhatted') quantities by

$$h_{ij} = \Phi^4 \hat{h}_{ij} \qquad (6.40)$$

$$K_{ij} = \Phi^{-2} \hat{K}_{ij}. \qquad (6.41)$$

These will satisfy the constraints (6.30), (6.31)!

6.5.5 Explicit time-symmetric data

Before we say a little more about maximal data, we wish to present some of the most popular examples for time-symmetric data some of which are also extensively used in numerical simulations. Hopefully these examples let you gain some intuition into the geometries and topologies involved and also let you anticipate the richness that a variable space structure gives to the solution space of one of the simplest equations in physics: the Laplace equation.

Restricting the solution strategy, outlined earlier, to the time-symmetric case, one first observes that for $K_{ij} = 0$ one has $\hat{K}_{ij} = 0$. The momentum constraint (6.31) is automatically satisfied and all that remains is equation (6.39), which now simply becomes the Laplace equation for the single scalar function Φ on the Riemannian manifold $(\Sigma, \hat{h}_{ij})$.

We now make a further simplifying assumption, namely that $\hat{h}_{ij}$ is, in fact, the *flat* metric. This will restrict our solution h_{ij} to a *conformally flat* geometry. It is not obvious how severe the loss of physically interesting solutions is by restricting ourselves to conformally flat metrics. But we will see that the latter already contain many interesting and relevant examples.

So let us solve Laplace's equation in flat space! Remember that Φ must be positive and approach one at spatial infinity (asymptotic flatness). We cannot

take $\Sigma = \mathbb{R}^3$ since the only solution to the Laplace equation in $\mathbb{R}^3$ which asymptotically approaches one is identically one. We must allow Φ to blow up at some points, which we can then remove from the manifold. In this way we let the solution tell us what topology to choose in order to have an everywhere regular solution. You might think that just removing singular points would be rather cheating, since the resulting manifold may turn out to be incomplete, that is, can be hit by a curve after finite proper length (you can go 'there'), even though Φ and hence the physical metric blows up at this point. If this were the case, one would definitely have to say what a solution on the completion would be. But, as a matter of fact, this cannot happen and the punctured space will turn out to be complete in terms of the physical metric.

6.5.5.1   One black hole

The simplest solution with one puncture (at $r = 0$) is just

$$\Phi(r, \theta, \varphi) = 1 + \frac{a}{r} \tag{6.42}$$

where a is a constant which we soon interpret and which must be positive in order for Φ to be positive everywhere. We cannot have other multipole contributions since they inevitably would force Φ to be negative somewhere. What is the geometry of this solution? The physical metric is

$$ds^2 = \left(1 + \frac{a}{r}\right)^4 (dr^2 + r^2\, d\theta^2 + \sin^2\theta\, d\varphi^2) \tag{6.43}$$

which is easily checked to be invariant under the inversion transformation on the sphere $r = a$:

$$r \rightarrow \frac{a^2}{r} \qquad \theta \rightarrow \theta \qquad \varphi \rightarrow \varphi. \tag{6.44}$$

This means that the region $r > a$ just looks like the region $r < a$ and that the sphere $r = a$ has the smallest area among all spheres of constant radius. It is a minimal surface, in fact even a totally geodesic submanifold, since it is the fixed point set of the isometry (6.44). Hence it is an apparent horizon, whose area follows from (6.43):

$$A = 16\pi(2a)^2. \tag{6.45}$$

Our manifold thus corresponds to a black hole (figure 6.1). Its mass can easily be computed from (6.35); one finds $m = 2a$. This manifold has two ends, one for $r \rightarrow \infty$ and one for $r \rightarrow 0$. They have the same geometry and hence the same ADM mass, as must be the case since individual and total mass clearly coincide for a single hole.

The data just written down correspond to the 'middle' slice right across the Kruskal (maximally extended Schwarzschild) manifold. Also, (6.43) is just the spatial part of the Schwarzschild metric in isotropic coordinates. Hence we know
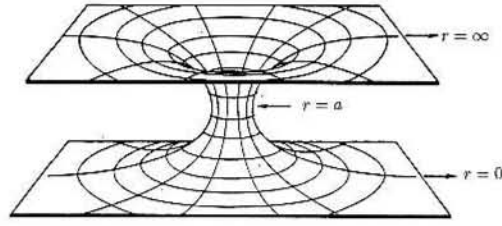
Figure 6.1. One black hole.

its entire future development in analytic form. Already for two holes this is no longer the case. Even the simplest two-body problem—head-on collision—has not been solved analytically in GR.

6.5.5.2 Two black holes

There is an obvious generalization of (6.42) by allowing two 'monopoles' of strength a_1 and a_2 at the punctures $\vec{x} = \vec{x}_1$ and $\vec{x} = \vec{x}_2$ respectively. The three-metric then reads:

$$ds^2 = \left(1 + \frac{a_1}{|\vec{x} - \vec{x}_1|} + \frac{a_2}{|\vec{x} - \vec{x}_2|}\right)^4 (dr^2 + r^2\,d\theta^2 + \sin^2\theta\,d\varphi^2). \quad (6.46)$$

The manifold has now three asymptotically flat ends, one for $|\vec{x}| \to \infty$, where the overall ADM mass M is measured, and one each for $|\vec{x} - \vec{x}_{1,2}| \to 0$. To see the latter, it is best to write the metric (6.46) in spherical polar coordinates $(r_1, \theta_1, \varphi_1)$ centered at $\vec{x}_1$, and then introduce the inverted radial coordinate given by $\bar{r}_1 = a_1^2/r_1$. In the limit $\bar{r}_1 \to \infty$, the metric then takes the form

$$ds^2 = \left(1 + \frac{a_1(1 + a_2/r_{12})}{\bar{r}_1} + O((1/\bar{r}_1)^2)\right)^4 (d\bar{r}^2 + \bar{r}^2(d\theta_1^2 + \sin^2\theta_1 d\varphi_1^2))$$

$$(6.47)$$

where $r_{12} = |\vec{x}_1 - \vec{x}_2|$. This looks just like a one-hole metric (6.43). Hence, if the black holes are well separated (compared to their size), the two-hole geometry looks like that depicted in figure 6.2. By comparison with the one-hole metric, we can immediately write down the ADM masses corresponding to the three ends $r, \bar{r}_{1,2} \to \infty$ respectively:

$$M = 2(a_1 + a_2) \qquad m_{1,2} = 2a_{1,2}(1 + \chi_{1,2}) \qquad \text{where } \chi_{1,2} = \frac{a_{2,1}}{r_{12}}.$$

$$(6.48)$$

Momenta and angular momenta clearly vanish (moment of time symmetry). Still assuming well-separated holes, i.e. $\chi_i = a_i/r_{12} \ll 1$, we can calculate the
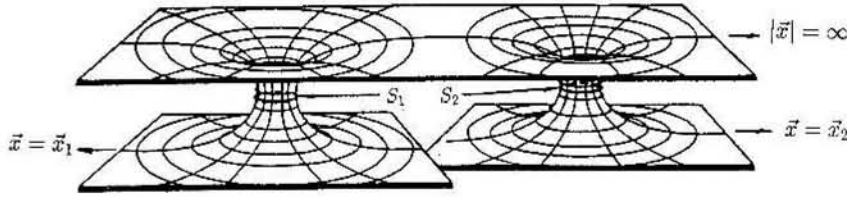
Figure 6.2. Two black holes well separated.

binding energy $\Delta E = M - m_1 - m_2$ as a function of the masses m_i and r_{12} and get

$$\Delta E = -\frac{m_1 m_2}{r_{12}} \left(1 - \frac{m_1 + m_2}{2r_{12}} + O((m_{1,2}/r_{12})^2) \right). \qquad (6.49)$$

The leading order is just the Newtonian expression for the binding energy of two point particles with masses $m_{1,2}$ at distance r_{12}. But there are corrections to this Newtonian form which tend to diminish the Newtonian value. Note also that (6.49) is still not in a good form since r_{12} is not an invariantly defined geometric distance measure. As such one might use the length ℓ of the shortest geodesic joining the two apparent horizons S_1 and S_2. Unfortunately these horizons are not easy to locate analytically and hence no closed form of $\ell(m_1, m_2, r_{12})$ exists which could be inverted to eliminate r_{12} in favour of ℓ.

Due to the difficulty of locating the two apparent horizons analytically, we also cannot write down an analytic expression for their area. But we can give upper and lower bounds as follows:

$$16\pi (2a_i)^2 < A_i < 16\pi [(2a_i(1 + \chi_i)]^2 = 16\pi m_i^2. \qquad (6.50)$$

The lower bound simply follows from the fact that the two-hole metric (6.46), if written down in terms of spherical polar coordinates about any of its punctures, equals the one-hole metric (6.43) plus a positive definite correction. The upper bound follows from the so-called Penrose inequality in Riemannian geometry (proven in [27]), which directly states that $16\pi m^2 \geq A$ for each asymptotically flat end, where m is the mass according to (6.35) and A is the area of the outermost (as seen from that end) minimal surface.

If the two holes approach each other to within a distance comparable to the sizes of the holes, the geometry changes in an essential way. This is shown in figure 6.3. The most important new feature is that new minimal surfaces form, in fact two [10], which both enclose the two holes. The outermost of these, as seen from the upper end, denoted by S_3 in figure 6.3, corresponds to the apparent horizon of the newly formed 'compound' black hole which contains the two old ones. For two black holes of equal mass, i.e. $a_1 = a_2 = a$, this happens approximately for a parameter ratio of $a/r_{12} = 0.65$ which, in an approximate
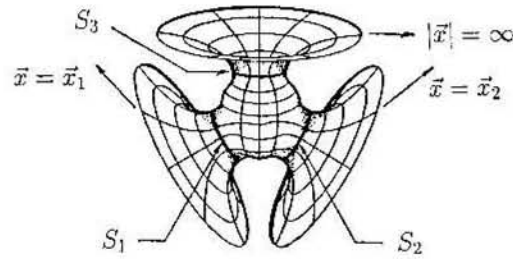
Figure 6.3. Two black holes after merging.

numerical translation into the ratio of individual hole mass to geodesic separation, reads $m/\ell \approx 0.26$.

6.5.5.3 *More than two black holes*

The method can be generalized in a straightforward manner to any number n of black holes with parameters $(a_i, \vec{x}_i)$, $i = 1, \ldots n$, for the punctures. The manifold Σ has now $n+1$ ends, one for $|\vec{x}| \to \infty$ and one for each $\vec{x} \to \vec{x}_i$. The expressions for the metric and masses are then given by the obvious generalizations of (6.46) and (6.48), respectively.

6.5.5.4 *Energy bounds from Hawking's area law*

Loosely speaking, Hawking's area law states that the surface of a black hole cannot decrease with time. (See [23] for a simple and complete outline of the traditional and technically slightly restricted version and [9] for the technically most complete proof known today.) Let us briefly explain this statement. If Σ is a Cauchy surface (a spacelike hypersurface in spacetime) and $\mathcal{H}$ the event horizon (a lightlike hypersurface in spacetime), the two intersect in a number of components (spacelike two-manifolds), each of which we assume to be a two-sphere. Each such two-sphere is called the surface of a black hole at time Σ. Let us pick one of them and call it B. Consider next a second Cauchy surface Σ' which lies to the future of Σ. The outgoing null rays of B intersect Σ' in a surface B', and the statement is now that the area of B' is larger than or equal to the area of B (to prove this one must assume the strong energy condition). Note that we deliberately left open the possibility that B' might be a proper subset of a black hole surface at time Σ', in case the original hole has merged in the meantime with another one. If this does not happen B' may be called the surface of the *same* black hole at the later time Σ.

Following an idea of Hawking's [24], this can be applied to the future evolution of multi black hole data as follows. As we have already mentioned, the event horizon lies on or outside the apparent horizon. Hence the area of

the 'surface' (as just defined) of a black hole is bounded below by the area of the corresponding apparent horizon, which in turn has the lower bound stated in (6.50). Suppose that, after a long time, our configuration settles into an approximately stationary state, at least for some interior region where gravitational radiation is no longer emitted. Since our data have zero linear and angular momentum, the final state is static and uniquely given by a single Schwarzschild hole of some final mass M_{final} and corresponding surface area $A_{\text{final}} = 16\pi M_{\text{final}}^2$. This is a direct consequence of known black hole uniqueness theorems (see [26] or [25, pp 157–86] for a summary). By the area theorem A_{final} is not less than the sum of all initial apparent horizon surface areas. This immediately gives

$$M_{\text{final}} \geq \sqrt{\sum_i A_i^{\text{initial}}/16\pi} \geq 2\sqrt{\sum_i^n a_i^2}. \tag{6.51}$$

In passing we remark that applied to a single black hole this argument shows that it cannot lose its mass below the value $m_{\text{ir}} := \sqrt{A^{\text{initial}}/16\pi}$, called its *irreducible mass*. Back to the multi-hole case, the total initial mass is given by the straightforward generalization of (6.48):

$$M_{\text{initial}} = \sum_i m_i = 2\sum_i a_i(1 + \chi_i) \qquad \text{where } \chi_i = \sum_{i \neq k} \frac{a_k}{|\vec{x}_i - \vec{x}_k|}. \tag{6.52}$$

By using these two equations, we can write down a lower bound for the fractional energy loss into gravitational radiation:

$$\frac{\Delta M}{M} := \frac{M_{\text{initial}} - M_{\text{final}}}{M_{\text{initial}}} \leq 1 - \frac{\sqrt{\sum_i a_i^2}}{\sum_i a_i(1 + \chi_i)}. \tag{6.53}$$

For a collision of n initially widely separated ($\chi_i \to 0$) holes of equal mass this becomes

$$\frac{\Delta M}{M} = 1 - \sqrt{1/n}. \tag{6.54}$$

For just two holes this means that at most 29% of their total rest mass can be radiated away. But this efficiency can be enhanced if the energy is distributed over a larger number of black holes.

Another way to raise the upper bound for the efficiency is to consider spinning black holes. For two holes the maximal value of 50% can be derived by starting with two extremal black holes (i.e. of maximal angular momentum: $J = m^2$ in geometric units) which merge to form a single unspinning black hole [24].

One can also envisage a situation where one hole participates in a scattering process but does not merge. Rather it gets kicked out of the collision zone and settles without spin (for simplicity) in a quasistationary state (for some time) far

apart. The question is what fraction of energy the area theorem allows it to lose. Let this be the kth hole. Then $m_k^{\text{final}} \geq 2a_k = m_k^{\text{initial}}/(1 + \chi_k)$. Hence

$$\frac{m_k^{\text{initial}} - m_k^{\text{final}}}{m_k^{\text{initial}}} \leq \frac{\chi_k}{1 + \chi_k} < 1 \tag{6.55}$$

showing that an appreciable efficiency can only be obtained if the data are such that χ_k is not too close to zero. This means that the kth hole was originally not too far from the others. This seems an unlikely process. Hence it is difficult to extract energy from a single unspinning hole.

For a single spinning black hole the situation is again different. Spinning it down from an extreme state to zero angular momentum sets an upper bound for the efficiency from the area law of 29%. This follows easily from the following relation between mass, irreducible mass, and angular momentum for a Kerr black hole (see e.g., formula (33.60)) [32]:

$$m^2 = m_{\text{ir}}^2 + J^2/4m_{\text{ir}}^2. \tag{6.56}$$

Setting $J = m^2$, one solves for $m_{\text{ir}}/m = \sqrt{1/2}$; hence $(m - m_{\text{ir}})/m = 1 - \sqrt{1/2} \approx 0.29$. It can, moreover, be shown [8] that this limit can be (theoretically) realized by the Penrose process (compare with chapter 5).

Needless to say, realistic processes may be far less efficient than this theoretical bound from the area law alone indicates. Recent numerical studies of the head-on (i.e. zero angular momentum) collision of two equal-mass black holes give a radiated energy in units of the total energy of only 10^{-3} [3]. With angular momentum the efficiency is, of course, expected to be much better. Here recent numerical investigations give an estimate of 3×10^{-2} for an inspiralling of two equal-mass non-spinning black holes from the innermost stable circular orbit [2]: still a long way from the theoretical upper bound.

6.5.5.5 Other topologies

Other topologies can be found which support initial data with apparent horizons. For example, instead of the 'Schwarzschild' manifold with $n + 1$ ends for n black holes (figure 6.4), one can find one which has just two ends for any number ≥ 2 of holes and which has been termed the Einstein–Rosen manifold [29] (figure 6.5). The difference from the data already discussed does not primarily lie in the physics they represent. After all their different topologies are hidden behind event horizons for the outside observer, even though their interaction energies are slightly different [20, 22]. However, the point we wish to stress here is that such data can analytically and numerically be more convenient, despite the fact that the underlying manifold might seem *topologically* more complicated. The reason is that these data have more symmetries and that coordinate systems can be found for which these symmetries take simple analytic expressions. For example, in the Einstein–Rosen manifold the upper and lower ends are isometrically related
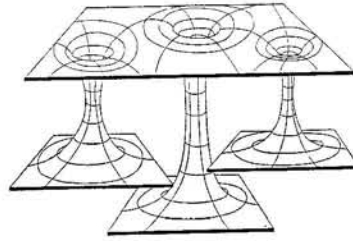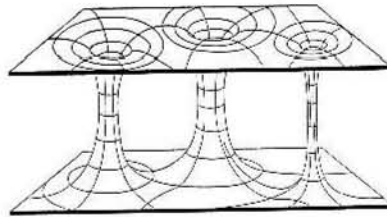
Figure 6.4. Multi-Schwarzschild.

Figure 6.5. Einstein–Rosen manifold.

by reflections about the minimal two-spheres in each connecting tube, with the fixed-point sets being the apparent horizons. Hence all the apparent horizons can be easily located analytically in the multi-hole Einstein–Rosen manifold, in contrast to the multi-hole Schwarzschild manifold.

Let us briefly explain this for two holes of equal mass. Here one starts again from $\mathbb{R}^3$ coordinatized by spherical bipolar coordinates. These are obtained from bipolar coordinates (μ, η) in the xz-plane by adding an azimuthal angle φ corresponding to a rotation about the z-axis, just as ordinary spherical polar coordinates are produced from ordinary polar coordinates in the xz-plane. The coordinates (μ, η) parametrize the xz-plane according to $\exp(\mu - i\eta) = (\xi + c)/(\xi - c)$, where $\xi = z + ix$ and $c > 0$ is a constant (figure 6.6). The lines of constant μ intersect those of constant η orthogonally. Both families consist of circles; those in the first family are centered on the z-axis with radii $c/\sinh \mu$ at $z = c \coth \mu$, and those in the second family on the x-axis with radii $c/|\sin \eta|$ at $|x| = c \cot \eta$.

Following an idea of Misner's [31], one can borrow the method of images from electrostatics (see, e.g., chapter 2.1 in [28]) in order to construct solutions Φ to the Laplace equation such that the metric $h_{ij} = \Phi^4 \delta_{ij}$ has a number of reflection isometries about two-spheres, one for each hole. In the two-hole case, one uses the two two-spheres $\mu = \pm\mu_0$ for some $\mu_0 > 0$, which then become the apparent horizons. By using these isometries, we can take two copies of our
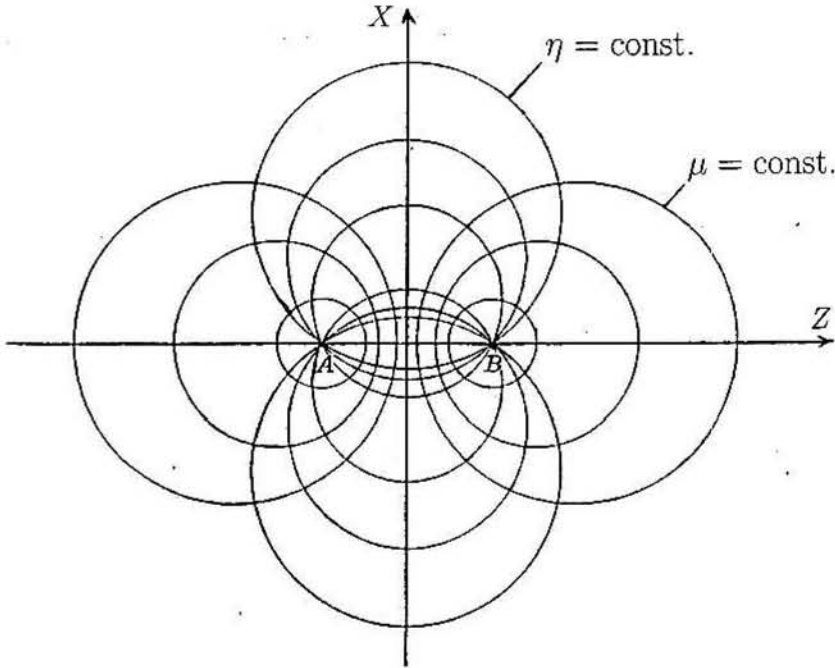
Figure 6.6. Bipolar coordinates.

initial manifold, excise the balls $|\mu| > \mu_0$ and glue the two remaining parts 'back to back' along the two boundaries $\mu = \mu_0$ and $\mu = -\mu_0$. The isometry property is necessary so that the metric continues to be smooth across the seam. This gives an Einstein–Rosen manifold with two tubes (or 'bridges', as they are sometimes called) connecting two asymptotically flat regions.

In fact, we could have just taken *one* copy of the original manifold, excised the balls $|\mu| > \mu_0$, and *mutually* glued together the two boundaries $\mu = \pm\mu_0$. This also gives a smooth metric across the seam and results in a manifold known as the Misner wormhole [30] (figure 6.7). Metrically the Misner wormhole is locally isometric to the Einstein–Rosen manifold with two tubes (which is its 'double cover'), but their topologies obviously differ. This means that for the observer outside the apparent horizons, these two data sets are indistinguishable. This is not quite true for the Einstein–Rosen and Schwarzschild data, which are not locally isometric. Even without exploring the region inside the horizons (which anyway is rendered impossible by existing results on topological censorship [16]) they slightly differ in their interaction energy and other geometric quantities, e.g. the tidal deformation of the apparent horizons.

The two parameters c and μ_0 now label the two-hole configurations of equal mass. (In the Schwarzschild case the two independent parameters were
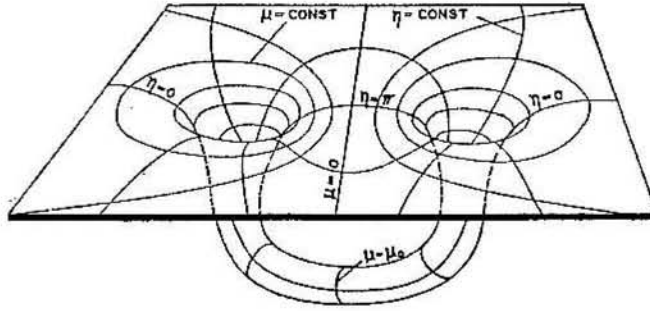
Figure 6.7. The Misner wormhole representing two black holes.

$a \equiv a_1 = a_2$ and $r_{12} \equiv |\vec{x}_1 - \vec{x}_2|$.) But unlike the Schwarzschild case, we can now give closed analytic expressions not only for the total mass M and individual mass m in terms of the two parameters, but also for the geodesic distance of the apparent horizons ℓ. (ℓ is used as the definition for the 'instantaneous distance of the two black holes'; for the Misner wormhole, where the two apparent horizons are identified, this corresponds to the length of the shortest geodesic winding once around the wormhole.) These read:

$$M = 4c \sum_{n=1}^{\infty} \frac{1}{\sinh n\mu_0} \qquad m = 2c \sum_{n=1}^{\infty} \frac{n}{\sinh n\mu_0} \qquad \ell = 2c(1 + 2m\mu_0). \quad (6.57)$$

You might rightly wonder what 'individual mass' should be if there is no internal end associated with each black hole where the ADM formula (6.35) can be applied. The answer is that there are alternative definitions of 'quasi-local mass' which can be applied even without asymptotic ends. The one we previously used for the expression of m is due to Lindquist [29] and is easy to compute in connection with the method of images but it lacks a deeper mathematical foundation. An alternative which is mathematically better founded is due to Penrose [34], which, however, is much harder to calculate and only applies to a limited set of situations (it agrees with the ADM mass whenever both definitions apply). Amongst them are, however, all time-symmetric conformally flat data, and for data above the Penrose mass has fortunately been calculated in [36]. The expression for m is rather complicated and differs from that given here. The difference is only of sixth order in an expansion in (mass/distance), though [20].

In summary, we see that the problem of setting up initial data for two black holes of given individual mass and given separation has no unique answer. Metrically as well as topologically different data sets can be found which have the same right to be called a realization of such a configuration. For holes without associated asymptotically flat ends no unambiguous definition for a quasi-local

mass exists.

6.5.6 Non-time-symmetric data

According to a prescription found by Bowen and York [35], we can add linear and angular momentum within the setting of maximal data. We can still use *conformally flat* data, i.e. set $\hat{h}_{ij} = \delta_{ij}$, on multiply punctured $\mathbb{R}^3$. Then the following two expressions add linear momentum P^i and spin angular momentum S^i to the puncture $\vec{x} = \vec{0}$:

$$\hat{K}_P^{ij} = \frac{3}{2r^2}(P^i n^j + P^j n^i - (\delta^{ij} - n^i n^j)(\vec{P} \cdot \vec{n})) \qquad (6.58)$$

$$\hat{K}_S^{ij} = \frac{3}{r^3}((\vec{S} \times \vec{n})^i n^j + (\vec{S} \times \vec{n})^j n^i). \qquad (6.59)$$

It is straightforward to check that these expressions satisfy (6.38) (note that $\hat{D}_i = \partial_i$). One can also check that these data will, indeed, give the proposed momenta and angular momenta at infinity (i.e. at the end $r \to \infty$). For this one may just use the 'hatted' quantities in (6.36) and (6.37), since the rescaling (6.41) does not influence the leading-order parts in the $1/r$ expansion of K, which alone contribute to these integrals. Linearity of all these equations in K allows us to just add K_P and K_S and get initial data for one black hole with given momentum $\vec{P}$ and (spin) angular momentum $\vec{S}$. Moreover, we can add any finite number of expressions of the kind (6.58) and (6.59) with parameters $\vec{P}_i$, $\vec{S}_i$ based at the puncture $\vec{x}_i$, where $i = 1, \ldots, n$. This then leads to a data set whose total linear and angular momentum is given by the sum $\sum_i \vec{P}_i$ and $\sum_i \vec{S}_i$ respectively. But one may not immediately conclude that the $\vec{P}_i$ and $\vec{S}_i$ are linear and angular momenta of the individual black holes. Rather, the latter must be calculated for the internal ends of the manifold and for this one needs to know Φ. The task then remains to solve (6.39) for the conformal factor, with blow-ups being allowed at the given punctures.

One interesting idea to facilitate solving (6.39) is first to split off the singular part of Φ, which blows up at the punctures $\{\vec{x}_1, \ldots, \vec{x}_n\}$ as $1/|\vec{x} - \vec{x}_i|$, from the regular remainder [6] (compare also [5]). One writes

$$\Phi = \frac{1}{\alpha} + U \qquad \text{with } \frac{1}{\alpha} := \sum_{i=n}^{n} \frac{a_i}{|\vec{x} - \vec{x}_i|} \qquad (6.60)$$

where the $a_i > 0$ may be freely prescribed. Inserting this into (6.39) gives

$$\Delta U + \beta(1 + \alpha U)^{-7} = 0 \qquad \text{with } \beta = \tfrac{1}{8}\alpha^7 K^{ij} K_{ij}. \qquad (6.61)$$

The point is now the following: for $\vec{x} \to \vec{x}_i$ the function α tends to zero as $|\vec{x} - \vec{x}_i|$; hence β, too, tends to zero as $|\vec{x} - \vec{x}_i|$. This means that (6.61) has continuous coefficients *everywhere* in $\mathbb{R}^3$ since the $1/|\vec{x} - \vec{x}_i|^6$ singularity at $\vec{x}_i$

of the K-squared term is cancelled by multiplication with α^7. (Note that this relies on using the K from (6.58), (6.59), which possess no $1/r^n$ terms with $n > 3$.) This means that equation (6.61) for U can be solved on all of $\mathbb{R}^3$, without the need to excise the points $\{\vec{x}_1, \ldots, \vec{x}_n\}$ and therefore without the need to specify 'inner' boundary conditions for U; only the 'outer' boundary condition $U(r \to \infty) \to 1$ remains. This simplification seems particularly useful in numerical implementations (compare [6]).

The total mass of our configuration is $M = \sum_i 2a_i$. The individual masses are determined just as in section 6.5.5.2 by introducing the inverted radial coordinate $\bar{r}_i = a_i^2/r_i$ and reading off the coefficient from the $1/2\bar{r}_i$ term in the $\bar{r}_i \to \infty$ expansion. One easily gets

$$m_i = 2a_i(U(\vec{x}_i) + \chi_i) \tag{6.62}$$

with χ_i as in (6.52).

The linear and angular momenta at, say, the kth end can also be calculated by using inverted coordinates, given by $\bar{\vec{x}} = (\vec{x} - \vec{x}_k)a_k^2/r_k^2$. Expressed in these coordinates, the 'hatted' (unphysical) extrinsic curvature tensor is given by $J^i{}_k J^j{}_l \hat{K}_{ij}$ where $J^i{}_k := \frac{\partial x^i}{\partial \bar{x}^k} = (a_k^2/\bar{r}^2)R^i_k$ with $R^i_k = \delta^i_k - 2n^i n_k$, which is an orthogonal matrix. The 'physical' extrinsic curvature is then obtained by multiplication with Φ^{-2} (compare with (6.41)). Now,

$$\Phi(\bar{\vec{x}}) = \frac{\bar{r}}{a_k}\left(1 + \frac{m_k}{2\bar{r}} + O((1/\bar{r})^2)\right)$$

so that

$$\bar{K}_{ij} = \left\{\left(\frac{a_k}{\bar{r}}\right)^6 + \text{terms} \propto \left(\frac{a_k}{\bar{r}}\right)^p\right\}(R^k_i R^l_j \hat{K}_{kl}) \qquad \text{where } p \geq 7. \tag{6.63}$$

Inserting the expression (6.58) for $\hat{K}_P$ results in a $1/\bar{r}^4$ fall-off so that the individual linear momenta are all zero *as measured from the internal ends*. One may say that the asymptotically flat internal ends represent the local rest frames of the black holes. Note that these rest frames are inertial since each black hole is freely falling. Inserting K_S from (6.59) gives a $1/\bar{r}^3$ fall-off and an angular momentum which is just $-\vec{S}_k$ for the kth end. (Here one uses that R^i_k is orientation-reversing orthogonal, hence changing the sign of ε_{ijk}, and that $R^i_k(\vec{S} \times \vec{n})^k = -(\vec{S} \times \vec{n})^i$.)

6.6 Problems and recent developments

In this final section we draw attention to some of the current problems and developments, without claiming completeness.

(1) Given black hole data for n holes of fixed masses and mutual separations (whatever definitions one uses here), one would like to minimize these

data on the amount of outgoing radiation energy. Any excess over the minimal amount can be said to be 'already contained' initially. But so far no local (in time) criterion is known which quantifies the amount of gravitational radiation in an initial data set. First hints at the possibility that some (Newman–Penrose) conserved quantities could be useful here were discussed in [13].

(2) Restricting ourselves to spatially conformally flat metrics seems to be too narrow. It has been shown that there are no conformally flat spatial slices in Kerr spacetime which are axisymmetric and reduce to slices of constant Schwarzschild time in the limit of vanishing angular momentum [18]. Accordingly, Bowen–York data, even for a single black hole, contain excess gravitational radiation due to the relaxation of the individual holes to Kerr form [19]; see also [33] for an informal discussion of this and related problems. An alternative to the Bowen–York data, which describe two spinning black holes and which reduce to Kerr data if the mass of one hole goes to zero, have been discussed in [12].

(3) Even for the simplest two-hole data (Schwarzschild or Einstein–Rosen) it is not known whether the evolving spacetime will have a suitably smooth asymptotic structure at future-lightlike infinity (i.e. 'scri-plus'). As a consequence, we still do not know whether we can give a rigorous mathematical meaning to the notion of 'energy loss by gravitational radiation' in this case of the simplest head-on collision of two black holes! The difficult analytical problems involved are studied in the framework of the so-called 'conformal field equations', see [17] (in particular section 4) for a summary and references.

(4) We usually like to ask 'Newtonian' questions, like: given two black holes of individual masses $m_{1,2}$ and mutual separation ℓ, what is their binding energy? For such a question to make sense, we need good concepts of *quasi-local mass* and *distance*. But these are ambiguous concepts in GR. Different definitions of 'quasi-local mass' and 'distance' amount to differences in the calculated binding energies which can be a few 10^{-3} times total energy at closest encounter [20]. This is of the same order of magnitude as the total energy lost into gravitational radiation found in [3] for the head-on (i.e. zero angular momentum) collision of two black holes modelled with Misner data.

6.7   Appendix: equation (6.2) satisfies the energy principle

By the 'energy principle' we understand the property that *all* energy of the self-gravitating system serves as source for the gravitational field. In this appendix, we wish to prove that (6.2) indeed satisfies this principle. For the uniqueness argument see [21].

Given a matter distribution ρ immersed in its own gravitational potential ϕ, suppose we redistribute the matter within a bounded region of space by

actively dragging it along the flow lines of a vector field $\vec{\xi}$ which vanishes outside some bounded region. The rate of change, $\delta\rho$, of the matter distribution is then determined through $\delta\rho \, dV = -L_{\vec{\xi}}(\rho \, dV) = -\vec{\nabla} \cdot (\vec{\xi}\rho) \, dV$, where $L_{\vec{\xi}}$ is the Lie derivative with respect to $\vec{\xi}$ and dV is the standard spatial volume element. Note that the latter also needs to be differentiated along $\vec{\xi}$, resulting in $L_{\vec{\xi}} \, dV = \vec{\nabla} \cdot \vec{\xi}$. Hence we have $\delta\rho = -\vec{\nabla} \cdot (\vec{\xi}\rho)$. The rate of work done to the system during this process is

$$\delta A = \int_{\mathbb{R}^3} dV \, \rho\vec{\xi} \cdot \vec{\nabla}\phi = -\int_{\mathbb{R}^3} dV \, \phi\vec{\nabla} \cdot (\vec{\xi}\rho) = \int_{\mathbb{R}^3} dV \, \phi\delta\rho \qquad (6.64)$$

where the integration by parts does not lead to surface terms due to $\vec{\xi}$ vanishing outside a bounded region. Equation (6.64) is still completely general, that is, independent of the field equation for ϕ. The field equation comes in when we assume that the process of redistribution is carried out adiabatically, which means that at each stage during the process ϕ satisfies its field equation with the instantaneous matter distribution. Our claim will be proven if under the hypothesis that ϕ satisfies (6.2) we can show that $\delta A = c^2\delta M_{\mathrm{G}}$, where M_{G} is defined in (6.5) and represents the total gravitating energy according to the field equation. Setting $\sqrt{\phi/c^2} = \psi$ and using the more convenient equation (6.3), we have

$$\delta A = \frac{c^4}{2\pi G} \int_{\mathbb{R}^3} dV \, \psi^2\delta\left[\frac{\Delta\psi}{\psi}\right] = \frac{c^4}{2\pi G} \int_{\mathbb{R}^3} dV \, [\psi\Delta(\delta\psi) - (\Delta\psi)\delta\psi]$$

$$= \frac{c^4}{2\pi G} \lim_{r\to\infty} \int_{S^2(r)} d\sigma \, \vec{n} \cdot [\psi\vec{\nabla}(\delta\psi) - (\vec{\nabla}\psi)\delta\psi]. \qquad (6.65)$$

Now, the fall-off condition for $r \to \infty$ implies that $\vec{\nabla}\psi$ falls off as fast as $1/r^2$ and $\delta\psi$ as $1/r$. Hence the second term in the last line of (6.65) does not contribute so that we may reverse its sign. This leads to

$$\delta A = \frac{c^4}{2\pi G}\delta \lim_{r\to\infty} \int_{S^2(r)} d\sigma \, (\vec{n} \cdot \vec{\nabla}\psi)\psi = \frac{c^4}{4\pi G}\delta \lim_{r\to\infty} \int_{S^2(r)} d\sigma \, \vec{n} \cdot \vec{\nabla}\phi$$

$$= c^2\delta M_{\mathrm{G}} \qquad (6.66)$$

which proves the claim.

References

[1] Arnowitt R, Deser S and Misner C 1961 Coordinate invariance and energy expressions in general relativity *Phys. Rev.* **122** 997–1006

[2] Baker J *et al* 2001 Plunge waveforms from inspiralling binary black holes *Phys. Rev. Lett.* **87** 121103 (gr-qc/0102037)

[3] Baker J *et al* 2000 Gravitational waves from black hole collisions via an eclectic approach *Class. Quant. Grav.* **17** L149–L56

[4] Beig R and Ó Murchadha N 1987 The Poincaré group as the symmetry group of canonical general relativity *Ann. Phys., NY* **174** 463–98

[5] Beig R 2000 Generalized Bowen–York initial data *Mathematical and Quantum Aspects of Relativity and Cosmology (Lecture Notes In Physics 537)* ed S Cotsakis and G Gibbons (Berlin: Springer) (gr-qc/0005043)

[6] Brandt S and Brügmann B 1997 A simple construction of initial data for multiple black holes *Phys. Rev. Lett.* **78** 3606–9 (gr-qc/9703066)

[7] Choquet-Bruhat Y and Geroch R 1969 Global aspects of the Cauchy problem in general relativity *Comm. Math. Phys.* **3** 334–57

[8] Christodoulou D 1970 Reversible and irreversible transformations on black-hole physics *Phys. Rev. Lett.* **25** 1596–7

[9] Chruściel P *et al* 2001 The area theorem *Ann. Henri Poincaré* **2** 109–78 (gr-qc/0001003)

[10] Čadež A 1974 Apparent horizons in the two-black hole problem *Ann. Phys., NY* **83** 449–57

[11] Dadhich N and Narlikar J (eds) 1998 *Gravitation and Relativity: at the Turn of the Millenium (Proc. GR-15 Conf. IUCAA Pune, India, December 1997)* (IUCAA)

[12] Dain S 2001 Initial data for two Kerr-like black holes *Phys. Rev. Lett.* **87** 121102 (gr-qc/0012023)

[13] Dain S and Valiente-Kroon J A 2002 Conserved quantities in a black hole collision *Class. Quant. Grav.* **19** 811–16 (gr-qc/0105109)

[14] Damour T 1987 The problem of motion in Newtonian and Einsteinian gravity *300 Years of Gravitation* ed S Hawking and W Israel (Cambridge: Cambridge University Press) pp 128–98

[15] Ehlers J *et al* 1976 Comments on gravitational radiation damping and energy loss in binary systems *Astrophys. J.* **208** L77–L81

[16] Friedman J *et al* 1993 Topological censorship *Phys. Rev. Lett.* **71** 1486–9
——1995 *Phys. Rev. Lett.* **75** 1872 (erratum)

[17] Friedrich H 1998 Einstein's equation and geometric asymptotics, see [11] 153–76 (gr-qc/9804009)

[18] Garat A and Price R 2000 Nonexistence of conformally flat slices of the Kerr spacetime *Phys. Rev.* D **61** 124011 (gr-qc/0002013)

[19] Gleiser R *et al* 1998 Evolving the Bowen-York initial data for spinning black holes *Phys. Rev.* D **57** 3401–7 (gr-qc/9710096)

[20] Giulini D 1990 Interaction energies for three-dimensional wormholes *Class. Quant. Grav.* **7** 1271–90

[21] Giulini D 1996 Consistently implementing the fields self-energy in Newtonian gravity *Phys. Lett.* A **232** 165–70 (gr-qc/9605011)

[22] Giulini D 1998 On the construction of time-symmetric black hole initial data, see [25], pp 224–43

[23] Giulini D 1998 Is there a general area theorem for black holes? *J. Math. Phys.* **39** 6603–6

[24] Hawking S 1971 Gravitational radiation from colliding black holes *Phys. Rev. Lett.* **26** 1344–6

[25] Hehl F *et al* (eds) 1998 *Black Holes: Theory and Observation (Lecture Notes in Physics 514)* (Berlin: Springer)

[26] Heusler M 1996 *Black Hole Uniqueness Theorems (Cambridge Lecture Notes in Physics)* (Cambridge: Cambridge University Press)

[27] Huisken G and Ilmanen T 1997 The Riemannian Penrose inequality *Int. Math. Res. Not.* **20** 1045–58

[28] Jackson J D 1975 *Classical Electrodynamics* (New York: Wiley) 2nd edn

[29] Lindquist R W 1963 Initial-value problem on Einstein–Rosen manifolds *J. Math. Phys.* **4** 938–50

[30] Misner C 1960 Wormhole initial conditions *Phys. Rev.* **118** 1110–11

[31] Misner C 1963 The method of images in geometrostatics *Ann. Phys., NY* **24** 102–7

[32] Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (New York: Freeman)

[33] Pullin J 1998 Colliding black holes: analytic insights, see [11] pp 87–105 (gr-qc/9803005)

[34] Penrose R 1982 Quasi-local mass and angular momentum in general relativity *Proc. R. Soc.* A **381** 53–63

[35] York J 1979 Kinematics and dynamics of general relativity *Sources of Gravitational Radiation* ed L Smarr (Cambridge: Cambridge University Press) pp 83–126

[36] Tod K P 1983 Some examples of Penrose's quasi-local mass construction *Proc. R. Soc.* A **388** 457–77

Chapter 7

Quantum aspects of black holes

Claus Kiefer
University of Cologne, Germany

In this chapter, a brief introduction is given to the quantum aspects of black holes. It is an important fact that black holes obey laws analogous to the laws of thermodynamics. Taking quantum theory into account, it becomes clear that black holes do in fact emit thermal radiation ('Hawking radiation') and possess an entropy. The physical meaning of Hawking temperature and black hole entropy is discussed in detail. These effects can only be observed if primordial black holes (relics from the early Universe) exist. The chapter therefore ends with a brief review of the current observational constraints on their existence.

7.1 Introduction

At the most fundamental level, black holes are genuine quantum objects. This holds irrespectively of the fact that direct quantum effects can only be observed for small black holes—black holes that cannot be formed by stellar collapse. For this reason the discussion in this chapter will be more theoretical in nature. But even a black hole as gigantic as the Galactic black hole will in the far future (if the Universe does not recollapse) be dominated by quantum effects and eventually evaporate. It is, however, possible that small black holes have been created in the very early Universe. For such primordial black holes quantum effects can be of direct observational significance in the present Universe. I shall thus devote my last section to a brief discussion of their relevance. In the first three sections I shall, however, give an introduction to the key theoretical developments— black hole mechanics, Hawking radiation, and the interpretation of the black hole entropy.

In my discussion I shall draw heavily from my review article (Kiefer 1999) where many technical details can be found. Other general references include the comprehensive book by Frolov and Novikov (1998), Wald (2001), Hehl *et*

al (1998), as well as the article by Bekenstein (1980) and the book by Thorne (1994).

7.2   The laws of black hole mechanics

It is a most amazing fact that black holes obey *uniqueness theorems* (Heusler 1996). If an object collapses to form a black hole, a stationary state is reached asymptotically. One can prove within Einstein–Maxwell theory that stationary black holes are uniquely characterized by only three parameters: mass M, angular momentum $J \equiv Ma$, and electric charge q. In this sense, black holes are objects much simpler than ordinary stars—given these parameters, they all look the same. All other degrees of freedom that might have been initially present have thus been radiated away, e.g. in the form of electromagnetic or gravitational radiation, during the collapse. Since the latter constitute some form of 'hair', one refers to the content of these theorems as *black holes have no hair*. The three parameters are associated with conservation laws at spatial infinity. In principle, one can thus decide about the nature of a black hole far away from the hole itself, without having to approach it. In astrophysical situations, the two parameters M and J suffice, since a charged object would rapidly discharge. The corresponding solution of Einstein's equations is called the Kerr solution (Kerr–Newman in the presence of charge). Stationary black holes are axially symmetric, with spherical symmetry being obtained as a special case for $J = 0$.

In the presence of other fields, the uniqueness theorems do not always hold, see, for example, Núñez *et al* (1998). This is, in particular, the case in the presence of non-Abelian gauge fields. In addition to charges at infinity, such 'coloured black holes' have to be characterized by additional variables, and it is necessary to approach the hole to determine them. The physical reason for the occurrence of such solutions is the nonlinear character of these gauge fields. Fields in regions closer to the black hole (that would otherwise be swallowed by the hole) are tied to fields far away from the hole (that would otherwise be radiated away) to reach an equilibrium situation. In most examples this equilibrium is, however, unstable and the corresponding black hole solution does not represent a physical solution. Since classical non-Abelian fields have never been observed (the description of objects such as quarks necessarily needs quantized gauge fields which, due to confinement, have no macroscopic limits), they will not be taken into account in the subsequent discussion.

In 1971, Stephen Hawking could prove an important theorem about stationary black holes—that their area can never decrease with time. More precisely, he showed that

> For a predictable black hole satisfying $R_{ab}k^ak^b \geq 0$ for all null k^a, the surface area of the *future* event horizon *never* decreases with time.

A 'predictable' black hole is one for which the cosmic censorship hypothesis holds—this is thus a major assumption for the area law. Cosmic censorship

states that all black holes occurring in nature have an event horizon, so that the singularity cannot be observed for far-away observers (the singularity is not 'naked'). I emphasize that the time asymmetry in this theorem comes into play because a statement is made about the future horizon, not the past horizon; an analogous statement for white holes would then be that the past event horizon never increases. I also emphasize that the area law only holds in classical theory, not in quantum theory (see section 7.3).

The area law seems to exhibit a close formal analogy to the Second Law of Thermodynamics—there the *entropy* can *never* decrease with time (for a closed system). However, the conceptual difference could not be more pronounced: while the Second Law is related to statistical behaviour, the area law is just a theorem in differential geometry. That the area law is in fact directly related to the Second Law will become clear in the course of this section.

Further support for this analogy is given by the existence of analogies to the other laws of thermodynamics. The Zeroth Law states that there is a quantity, the temperature, that is constant on a body in thermal equilibrium. Does there exist an analogous quantity for a black hole? One can in fact prove that the surface gravity κ is constant over the event horizon (Wald 1984). For a Kerr black hole, κ is given by

$$\kappa = \frac{\sqrt{(GM)^2 - a^2}}{2GMr_+} \xrightarrow{a \to 0} \frac{1}{4GM} = \frac{GM}{R_0^2} \tag{7.1}$$

where r_+ denotes the location of the event horizon. One recognizes in the Schwarzschild limit the well-known expression for the Newtonian gravitational acceleration. (There $R_0 \equiv 2GM$ denotes the Schwarzschild radius.) One can show for a static black hole that κ is the limiting force that must be exerted at infinity to hold a unit test mass in place when approaching the horizon. This justifies the name surface gravity.

With a tentative formal relation between surface gravity and temperature, and between area and entropy, the question arises as to whether a First Law of thermodynamics can be proved. This can in fact be done and the result for a Kerr–Newman black hole is

$$dM = \frac{\kappa}{8\pi G} dA + \Omega_H dJ + \Phi dq \tag{7.2}$$

where A, Ω_H, Φ denote the area of the event horizon, the angular velocity of the black hole, and the electrostatic potential, respectively. This relation can be obtained by conceptually different methods: a *physical process version* in which a stationary black hole is altered by infinitesimal physical processes, and an *equilibrium state version* in which the areas of two stationary black hole solutions to Einstein's equations are compared. Both methods lead to the same result (7.2).

Since M is the energy of the black hole, (7.2) is the analogue of the First Law of Thermodynamics given by

$$dE = T dS - p dV + \mu dN. \tag{7.3}$$

Table 7.1. Analogy between the laws of thermodynamics and the laws of black-hole mechanics.

Law	Thermodynamics	Stationary black holes
Zeroth	T constant on a body in thermal equilibrium	κ constant on the horizon of a black hole
First	$dE = T\,dS - p\,dV + \mu\,dN$	$dM = \frac{\kappa}{8\pi G}\,dA + \Omega_H\,dJ + \Phi\,dq$
Second	$dS \geq 0$	$dA \geq 0$
Third	$T = 0$ cannot be reached	$\kappa = 0$ cannot be reached

'Modern' derivations of (7.2) make use of both Hamiltonian and Lagrangian methods of general relativity. For example, the First Law follows from an arbitrary diffeomorphism invariant theory of gravity whose field equations can be derived from a Lagrangian.

What about the Third Law of Thermodynamics? A 'physical process version' was proved by Israel—it is impossible to reach $\kappa = 0$ in a finite number of steps, although it is unclear whether this is true under all circumstances (Farrugia and Hajicek 1979). This corresponds to the 'Nernst version' of the Third Law. The stronger 'Planck version', which states that the entropy goes to zero (or a material-dependent constant) if the temperature approaches zero, does not seem to hold. The analogies are summarized in table 7.1.

The identification of the horizon area with the entropy for a black hole can be obtained from a conceptually different point of view. If a box with, say, thermal radiation of entropy S is thrown into the black hole, it seems as if the Second Law could be violated, since the black hole is characterized only by mass, angular momentum, and charge, but nothing else. The demonstration that the Second Law is fulfilled leads immediately to the concept of a black hole entropy, as will now be discussed (Bekenstein 1980, Sexl and Urbantke 1983).

Consider a box with thermal radiation of mass m and temperature T lowered from a spaceship far away from a spherically symmetric black hole towards the hole (figure 7.1). As an idealization, both the rope and the walls are assumed to have negligible mass. At a coordinate distance r from the black hole, the energy of the box is given by

$$E_r = m\sqrt{1 - \frac{2GM}{r}} \xrightarrow{r \to R_0} 0. \tag{7.4}$$

If the box is lowered down to the horizon, the energy gain is thus given by m. The box is then opened and thermal radiation of mass δm escapes into the hole. If the box is then closed and brought back again to the spaceship, the energy loss is $m - \delta m$. In total the energy δm of the thermal radiation can be transformed into work with an efficiency of $\eta = 1$. This looks as if one possessed a perpetuum mobile of the second kind.
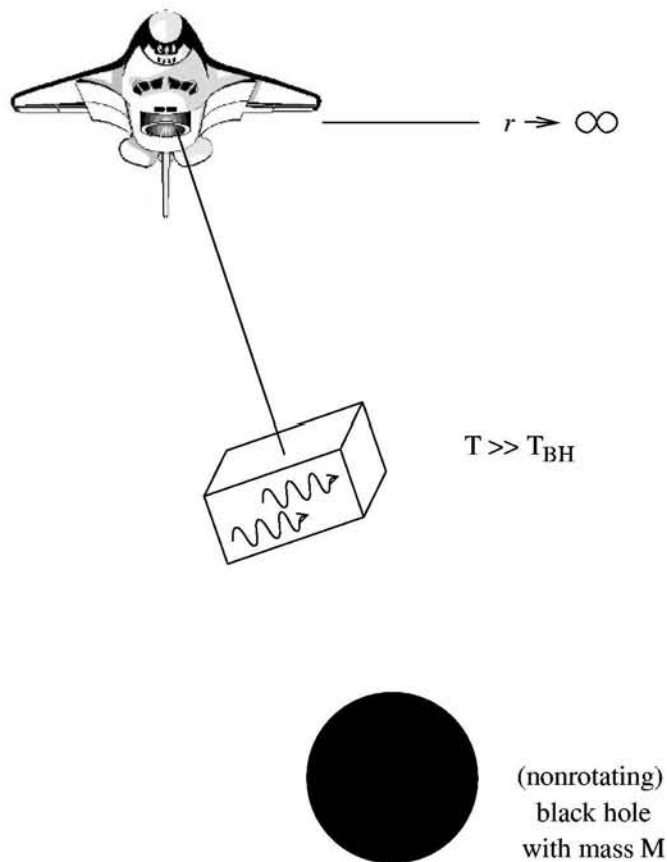
Figure 7.1. Thought experiment to demonstrate the Second Law of Thermodynamics for black holes.

The key to the resolution of this apparent paradox lies in the observation that the box must be big enough to contain the wavelength of the enclosed radiation. This, in turn, leads to a lower limit on the distance to which the box can approach the horizon. Therefore, only part of δm can be transformed into work, as I shall show now.

According to Wien's law, one must have a linear extension of the box of at least

$$\lambda_{\max} \approx \frac{\hbar}{k_{\mathrm{B}} T} \tag{7.5}$$

where k_{B} denotes Boltzmann's constant. I emphasize at this stage that Planck's constant $\hbar$ comes into play. The box can then be lowered down to the coordinate

distance δr (assumed to be $\ll 2GM$) from the black hole where, according to the Schwarzschild metric, the relation between δr and λ_{max} is

$$\lambda_{max} \approx \int_{2GM}^{2GM+\delta r} \left(1 - \frac{2GM}{r}\right)^{-\frac{1}{2}} \mathrm{d}r \approx 2\sqrt{2GM\delta r} \implies \delta r \approx \frac{\lambda_{max}^2}{8GM}.$$

According to (7.4), the energy of the box at $r = 2GM + \delta r$ is

$$E_{2GM+\delta r} = m\sqrt{1 - \frac{2GM}{2GM + \delta r}} \approx \frac{m\lambda_{max}}{4GM} \approx \frac{m\hbar}{4Gk_BTM}.$$

Recalling that, according to (7.2), the formal temperature of the black hole, T_{BH}, is proportional to the surface gravity $\kappa = 1/(4GM)$, the energy of the box before opening is

$$E_{2GM+\delta r}^{(before)} \approx m\frac{T_{BH}}{T}$$

while after opening it is

$$E_{2GM+\delta r}^{(after)} \approx (m - \delta m)\frac{T_{BH}}{T}.$$

The efficiency of transforming thermal radiation into work is thus given by

$$\eta \approx \left(\delta m - \delta m\frac{T_{BH}}{T}\right) \bigg/ \delta m = 1 - \frac{T_{BH}}{T} < 1$$

which is the well-known Carnot limit for the efficiency of heat engines. From the First Law (7.2) one then finds for the entropy of the black hole $S_{BH} \propto A = 16\pi(GM)^2$. It is this agreement of conceptually different approaches to black hole thermodynamics that gives us confidence in the physical meaning of these concepts. In the next section I shall show how all these formal results can be physically interpreted in the context of quantum theory.

7.3 Hawking radiation

We have already seen in the thought experiment discussed in the previous section that $\hbar$ enters the scene, see (7.5). That Planck's constant has to play a role can also be seen from the First Law (7.2). Since $T_{BH}\, \mathrm{d}S_{BH} = \kappa/(8\pi G)\, \mathrm{d}A$, one must have

$$T_{BH} = \frac{\kappa}{G\zeta} \qquad S_{BH} = \frac{\zeta A}{8\pi}$$

with an undetermined factor ζ. What is the dimension of ζ? Since S_{BH} has the dimension of Boltzmann's constant k_B, k_B/ζ must have the dimension of a length squared. There is, however, only one fundamental length available, the Planck length

$$l_P = \sqrt{G\hbar} \approx 10^{-33} \text{ cm.} \tag{7.6}$$
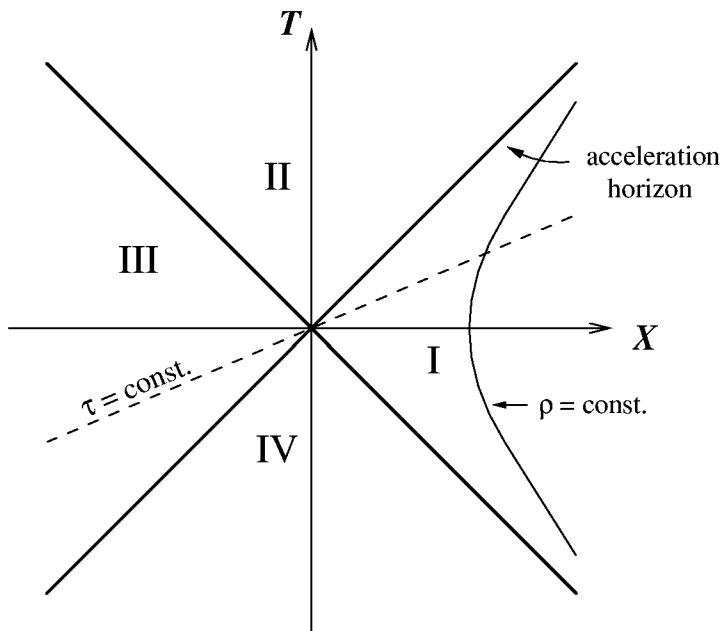
Figure 7.2. Uniformly accelerated observer in Minkowski space.

(For string theory, this may be replaced by the fundamental string length.) Therefore,

$$T_{BH} \propto \frac{\hbar\kappa}{k_B} \qquad S_{BH} \propto \frac{k_B A}{G\hbar}. \qquad (7.7)$$

The precise factors in (7.7) were determined in a pioneering paper by Hawking (1975). The key ingredient in his discussion is the behaviour of *quantum* fields on the background of an object collapsing to form a black hole. Similar to the situation of an external electric field (Schwinger effect), there is no uniquely defined notion of a *vacuum*. This leads to the occurrence of particle creation. The peculiarity of the black hole case is the *thermal* distribution of the particles created.

There exists an analogous effect already in flat spacetime, discussed by Unruh (1976), with earlier related work by Fulling (1973) and Davies (1975). In the following I shall briefly describe this effect.

Whereas all inertial observers in Minkowski space agree on the notion of vacuum (and therefore on particles), this no longer holds for *non-inertial* observers. Consider an observer who is uniformly accelerating along the X-direction in $(1 + 1)$-dimensional Minkowski spacetime (figure 7.2). The Minkowski Cartesian coordinates are labelled here by upper-case letters. The

orbit of this observer is the hyperbola shown in figure 7.2. One recognizes that, as in the Kruskal diagram for the Schwarzschild metric, the observer encounters a horizon (here an 'acceleration horizon'). There is, however, no singularity behind this horizon. Region I is a globally hyperbolic spacetime on its own—the so-called *Rindler spacetime*. This spacetime can be described by coordinates (τ, ρ) which are connected to the Cartesian coordinates via the coordinate transformation

$$\begin{pmatrix} T \\ X \end{pmatrix} = \rho \begin{pmatrix} \sinh a\tau \\ \cosh a\tau \end{pmatrix} \tag{7.8}$$

where a is a constant (the orbit in figure 7.2 describes an observer with acceleration a, who has $\rho = 1/a$).
Since

$$ds^2 = dT^2 - dX^2 = a^2\rho^2\,d\tau^2 - d\rho^2 \tag{7.9}$$

the orbits $\rho = $ constant are also orbits of a timelike Killing field $\partial/\partial\tau$. It is clear that τ corresponds to the external Schwarzschild coordinate t and that ρ corresponds to r. As in the Kruskal case, $\partial/\partial\tau$ becomes spacelike in regions II and IV.

The analogy with Kruskal becomes even more transparent if the Schwarzschild metric is expanded around the horizon at $r = 2GM$. Introducing $\rho^2/(8GM) = r - 2GM$ and recalling (7.1), one has

$$ds^2 \approx \kappa^2\rho^2\,dt^2 - d\rho^2 - \frac{1}{4\kappa^2}\,d\Omega^2. \tag{7.10}$$

Comparison with (7.9) shows that the first two terms on the right-hand side of (7.10) correspond exactly to the Rindler spacetime (7.9) with the acceleration a replaced by the surface gravity κ. The last term[1] in (7.10) describes a two-sphere with radius $(2\kappa)^{-1}$.

How does the accelerating observer experience the standard Minkowski vacuum $|0\rangle_M$? The key point is that the vacuum is a *global* state correlating regions I and III in figure 7.2 (similar to Einstein–Podolsky–Rosen correlations), but that the accelerated observer is restricted to region I. Considering, for simplicity, the case of a massless scalar field, the global vacuum state comprising the regions I and II can be written in the form

$$|0\rangle_M = \prod_\omega \sqrt{1 - e^{-2\pi\omega a^{-1}}} \sum_n e^{-n\pi\omega a^{-1}} |n_\omega^I\rangle \otimes |n_\omega^{II}\rangle \tag{7.11}$$

where $|n_\omega^I\rangle$ and $|n_\omega^{II}\rangle$ are n-particle states with frequency $\omega = |k|$ in regions I and II, respectively. Expression (7.11) is an example of the Schmidt expansion of two

---

[1] It is this term that is responsible for the non-vanishing curvature of (7.10) compared to the flat-space metric (7.9) whose extension into the (neglected) other dimensions would be just $-dY^2 - dZ^2$.

entangled quantum systems, see, e.g., Giulini *et al* (1996); note also the analogy of (7.11) with a BCS state in the theory of superconductivity.

For an observer restricted to region I, state (7.11) cannot be distinguished, by operators with support in I only, from a density matrix that is found from (7.11) by tracing out all degrees of freedom in region II,

$$\rho_{\mathrm{I}} \equiv \mathrm{Tr}_{\mathrm{II}} |0\rangle_M \langle 0|_M$$

$$= \prod_\omega (1 - e^{-2\pi\omega a^{-1}}) \sum_n e^{-2\pi n\omega a^{-1}} |n_\omega^{\mathrm{I}}\rangle \langle n_\omega^{\mathrm{I}}|. \tag{7.12}$$

Note that the density matrix ρ_{I} has exactly the form corresponding to a thermal canonical ensemble with temperature

$$T_{\mathrm{U}} = \frac{\hbar a}{2\pi k_{\mathrm{B}}} \approx 4 \times 10^{-23} a \left[\frac{\mathrm{cm}}{\mathrm{s}^2}\right] \mathrm{K}. \tag{7.13}$$

An observer who is accelerating uniformly through Minkowski space thus sees a *thermal* distribution of particles. This is an important manifestation of the non-uniqueness of the vacuum state in quantum field theory, even for flat spacetime. A more detailed discussion invoking models of particle detectors confirms this result.

The 'Unruh temperature' (7.13), although being very small for most accelerations, might be observable for electrons in storage rings where spin precession is used as a 'detector' (Leinaas 2001). Due to the circular nature of the accelerator, the spectrum of the observed particles is not then thermal. Since this would complicate the direct comparison with the Hawking effect, other proposals for measuring (7.13), for example by means of ultra-intense lasers (Chen and Tajima 1999), also exist.

I shall now turn to the case of black holes. From the form of the line element near the horizon, (7.10), one can already anticipate that—according to the equivalence principle—a black hole radiates with a temperature as specified in (7.13) with a being replaced by κ. This is, in fact, what Hawking (1975) found. The temperature reads:

$$T_{\mathrm{BH}} = \frac{\hbar \kappa}{2\pi k_{\mathrm{B}}}. \tag{7.14}$$

For the total luminosity of the black hole one finds

$$L = -\frac{\mathrm{d}M}{\mathrm{d}t} = \frac{1}{2\pi} \sum_{l=0}^{\infty} (2l+1) \int_0^\infty \mathrm{d}\omega\, \omega \frac{\Gamma_{\omega l}}{e^{2\omega\pi\kappa^{-1}} - 1}. \tag{7.15}$$

The term $\Gamma_{\omega l}$—called the 'grey body factor' because it encodes a deviation from the black body spectrum—takes into account the fact that some of the particle modes are backscattered into the black hole by means of spacetime curvature.

For the special case of the Schwarzschild metric where $\kappa = (4GM)^{-1}$, (7.14) becomes

$$T_{\mathrm{BH}} = \frac{\hbar}{8\pi Gk_{\mathrm{B}}M} \approx 6.2 \times 10^{-8} \frac{M_\odot}{M} \mathrm{K}. \tag{7.16}$$

For solar-mass black holes (and even more so for the Galactic black hole), this is of course utterly negligible—the black hole absorbs much more from the ubiquitous 3K-microwave background radiation than it radiates itself.

One can, however, estimate the lifetime of a black hole by making the plausible assumption that the decrease in mass is equal to the energy radiated to infinity. Using the Stefan–Boltzmann law, one gets

$$\frac{\mathrm{d}M}{\mathrm{d}t} \propto -AT_{\mathrm{BH}}^4 \propto -M^2 \times \left(\frac{1}{M}\right)^4 = -\frac{1}{M^2}$$

which, when integrated, yields

$$t(M) \propto (M_0^3 - M^3) \approx M_0^3. \tag{7.17}$$

Here M_0 is the initial mass. It has been assumed that after the evaporation $M \ll M_0$. Very roughly, the lifetime of a black hole is thus given by

$$\tau_{\mathrm{BH}} \approx \left(\frac{M_0}{m_{\mathrm{P}}}\right)^3 t_{\mathrm{P}} \approx 10^{65} \left(\frac{M_0}{M_\odot}\right)^3 \text{ years} \tag{7.18}$$

(m_{P} and t_{P} denote Planck mass and Planck time: $m_{\mathrm{P}} = \hbar/l_{\mathrm{P}}$, $t_{\mathrm{P}} = l_{\mathrm{P}}$). The Galactic black hole thus has a lifetime of about 3×10^{85} years! If in the early universe primordial black holes with $M_0 \approx 5 \times 10^{14}$ g were created, they would evaporate at the present age of the universe, see section 7.5.

A very detailed investigation into black hole evaporation was made by Page (1976). He found that for $M \gg 10^{17}$ g the power emitted from an (uncharged, non-rotating) black hole is

$$P \approx 2.28 \times 10^{-54} L_\odot \left(\frac{M}{M_\odot}\right)^{-2}$$

81.4% of which is in neutrinos (he considered only electron- and muon-neutrinos), 16.7% in photons, and 1.9% in gravitons, assuming that there are no other massless particles around. Since a black hole emits *all* existing particles in Nature, this result would be changed by the existence of massless supersymmetric or other particles. In the range 5×10^{14} g $\ll M \ll 10^{17}$ g, Page found

$$P \approx 6.3 \times 10^{16} \left(\frac{M}{10^{15} \text{ g}}\right)^{-2} \frac{\text{erg}}{\text{s}}$$

45% of which is in electrons and positrons, 45% in neutrinos, 9% in photons, and 1% in gravitons. Massive particles with mass m are only suppressed if $k_{\mathrm{B}}T_{\mathrm{BH}} < m$. For $M < 5 \times 10^{14}$ g higher-mass particles are also emitted.

All of these derivations use the approximation that the spacetime background remains classical[2]. In a theory of quantum gravity, however, such a picture cannot

---

[2] This limit is referred to as the semiclassical approximation to quantum gravity (see, e.g., Kiefer 1994).

be maintained. Since the black hole becomes hotter while radiating, see (7.16), its mass will eventually enter the quantum-gravity domain $M \approx m_P$, where the semiclassical approximation breaks down. The evaporation then enters the realm of speculation. As an intermediate step one might consider the heuristic 'semiclassical' Einstein equations,

$$R_{ab} - \tfrac{1}{2} g_{ab} R = 8\pi G \langle T_{ab} \rangle \qquad (7.19)$$

where on the right-hand side the quantum expectation value of the energy–momentum tensor appears. The evaluation of $\langle T_{ab} \rangle$—which requires regularization and renormalization—is a difficult subject on its own (Frolov and Novikov 1998). The renormalized $\langle T_{ab} \rangle$ is essentially unique (its ambiguities can be absorbed in coupling constants) if certain sensible requirements are imposed. Evaluating the components of the renormalized $\langle T_{ab} \rangle$ near the horizon, one finds that there is a flux of *negative energy* into the hole. Clearly this leads to a decrease of the black hole's mass. These negative energies represent a typical quantum effect and are well known from the—accurately measured—Casimir effect. This occurrence of negative energies is also responsible for the breakdown of the classical area law discussed in section 7.2.

The negative flux near the horizon also lies at the heart of the 'pictorial' representation of Hawking radiation that is often used, see, e.g., Parikh and Wilczek (2000). In vacuum, virtual pairs of particles are created and destroyed. However, close to the horizon, one partner of this virtual pair might fall into the black hole, thereby liberating the other partner to become a real particle and escaping to infinity as Hawking radiation. The global quantum field exhibits quantum entanglement between the inside and outside of the black hole, similar to the case of the accelerated observer discussed earlier.

I want to end this section by giving explicit expressions for the Hawking temperature (7.14) in the case of rotating and charged black holes. For the Kerr solution, one has

$$k_B T_{BH} = \frac{\hbar \kappa}{2\pi} = 2 \left(1 + \frac{M}{\sqrt{M^2 - a^2}} \right)^{-1} \frac{\hbar}{8\pi M} < \frac{\hbar}{8\pi M}. \qquad (7.20)$$

Rotation thus reduces the Hawking temperature. For the Reissner–Nordström solution (describing a charged spherically symmetric black hole) one has

$$k_B T_{BH} = \frac{\hbar}{8\pi M} \left(1 - \frac{(Gq)^4}{r_+^4} \right) < \frac{\hbar}{8\pi M}. \qquad (7.21)$$

Thus, electric charge also reduces the Hawking temperature. For an extremal black hole, $r_+ = GM = \sqrt{G}|q|$, and thus $T_{BH} = 0$.

7.4   Interpretation of entropy

We have seen in the previous section that—if quantum theory is taken into account—black holes emit thermal radiation with the temperature (7.14). Consequently, the laws of black hole mechanics discussed in section 7.2 do, indeed, have a physical interpretation as thermodynamical laws—black holes *are* thermodynamical systems.

From the First Law (7.2) one can therefore also infer an expression for the black hole entropy. From $dM = T_{BH}\,dS_{BH}$ one finds the 'Bekenstein–Hawking entropy':

$$S_{BH} = \frac{k_B A}{4G\hbar} \tag{7.22}$$

in which the unknown factor in (7.7) has now been fixed. For the special case of a Schwarzschild black hole, this yields

$$S_{BH} = \frac{k_B \pi R_0^2}{G\hbar}. \tag{7.23}$$

It can easily be estimated that S_{BH} is much bigger than the entropy of the star that collapsed to form the black hole. The entropy of the sun, for example, is $S_\odot \approx 10^{57} k_B$, whereas the entropy of a solar-mass black hole is about $10^{77} k_B$, which is 20 orders of magnitude larger! For the Galactic black hole, the entropy is $S_{GBH} \approx 10^{90} k_B$ which is 100 times the entropy of the Universe. (Under the 'entropy of the Universe' I understand the entropy of the present Universe up to the Hubble radius without taking black holes into account. It is dominated by the entropy of the cosmic microwave background radiation.)

Can a physical interpretation of this huge discrepancy be given? Up to now, the laws of black hole mechanics are only phenomenological thermodynamical laws. The central open question therefore is: can S_{BH} be derived from quantum-statistical considerations? This would mean that S_{BH} could be calculated from a Gibbs-type formula according to

$$S_{BH} \stackrel{?}{=} -k_B \operatorname{Tr}(\rho \ln \rho) \equiv S_{SM} \tag{7.24}$$

where ρ denotes an appropriate density matrix and S_{SM} is the quantum-statistical entropy; S_{BH} would then somehow correspond to the number of quantum microstates that are consistent with the macrostate of the black hole that is—according to the no-hair theorem—uniquely characterized by mass, angular momentum, and charge. Some important questions are:

- Does S_{BH} correspond to states hidden behind the horizon?
- Or does S_{BH} correspond to the number of possible initial states?
- What are the microscopic degrees of freedom?
- Where are they located (if at all)?
- Can one understand the universality of the result?

- What happens to S_{BH} after the black hole has evaporated?
- Is the entropy a 'one-loop' or a 'tree-level' effect?

The attempts to calculate S_{BH} by state counting are usually done in the 'one-loop limit' of quantum field theory in curved spacetime—this is the limit where gravity is classical but non-gravitational fields are fully quantum, and it is the limit where the Hawking radiation (7.14) has been derived. The expression (7.22) can already be calculated from the so-called 'tree level' of the theory, where only the gravitational degrees of freedom are taken into account. Usually a saddle-point approximation for a Euclidean path integral is being performed. Such derivations are, however, equivalent to derivations within classical thermodynamics, cf Wald (2001).

If the entropy (7.22) is to make sense, there should be a generalized Second Law of Thermodynamics according to which

$$\frac{d}{dt}(S_{BH} + S_M) \geq 0 \tag{7.25}$$

where S_M denotes the non-gravitational entropy. The validity of (7.25), although far from being proven in general, has been shown in a variety of thought experiments. One of the most instructive of these was devised by Unruh and Wald. It makes use of the box shown in figure 7.1 that is adiabatically lowered towards a (spherically symmetric) black hole.

At asymptotic infinity $r \to \infty$, the black hole radiation is given by (7.14). However, for finite r, the temperature is modified by the occurrence of a redshift factor $\chi(r) \equiv (1 - 2GM/r)^{1/2}$ in the denominator. Since the box is not in free fall, it is accelerated with an acceleration a. From the relation (Wald 1984)

$$\kappa = \lim_{r \to R_0} (a\chi) \tag{7.26}$$

one has

$$T_{BH}(r) = \frac{\hbar\kappa}{2\pi k_B \chi(r)} \xrightarrow{r \to R_0} \frac{\hbar a}{2\pi k_B} \tag{7.27}$$

which is just the Unruh temperature (7.13)! This means that a freely falling observer near the horizon observes no radiation at all, and the whole effect (7.27) comes from the observer (or box) being non-inertial with acceleration a.

The analysis of Unruh and Wald, which is a generalization of the thought experiment discussed at the end of section 7.2, shows that the entropy of the black hole increases at least by the entropy of the Unruh radiation displaced at the floating point—this is the point where the gravitational force (pointing downwards) and the buoyancy force from the Unruh radiation (7.27) are in equilibrium. Interestingly, it is just the application of 'Archimedes' principle' to this situation that rescues the generalized Second Law (7.25).

An inertial, i.e. freely falling, observer does not see any Unruh radiation. How does he/she interpret this result? For him/her the box is accelerated and

therefore the interior of the box fills up with negative energy and pressure—a typical quantum effect that occurs if a 'mirror' is accelerated through the vacuum. The 'floating point' is then reached after this negative energy is so large that the total energy of the box is zero.

I want to conclude this section with some speculations about the final stages of black hole evolution and the information-loss problem. The point is that— in the semiclassical approximation used by Hawking—the radiation of a black hole seems to be purely thermal. If the black hole evaporates completely and leaves only thermal radiation behind, one would have a conflict with established principles in quantum theory: any initial state (in particular a pure state) would evolve into a mixed state. In ordinary quantum theory, because of the unitary evolution of the total system, this cannot happen. Formally, $\mathrm{Tr}\,\rho^2$ remains *constant* under the von Neumann equation; the same is true for the entropy $S_{\mathrm{SM}} = -k_\mathrm{B}\,\mathrm{Tr}(\rho \ln \rho)$: For a unitarily evolving system, there is no increase in entropy. If these laws were violated during black hole evaporation, information would be destroyed. This is, indeed, the speculation that Hawking made after his discovery of black hole radiation. The attitudes towards this *information-loss problem* can be roughly divided into the following classes:

- The information is indeed lost during black hole evaporation, and the quantum-mechanical Liouville equation is replaced by an equation of the form

$$\rho \longrightarrow \$\rho \neq S\rho S^\dagger \qquad (7.28)$$

  where $\$$ is Hawking's dollar matrix which generalizes the ordinary S-matrix S.
- The full evolution is in fact unitary; the black hole radiation contains subtle quantum correlations that cannot be seen in the semiclassical approximation.
- The black hole does not evaporate completely, but leaves a 'remnant' with mass in the order of the Planck mass that carries the whole information.

In my opinion, the information-loss problem is only a pseudoproblem. Already in the original calculation of Hawking (1975) only pure states appear. Reference to thermal radiation is being made because the particle number operator in the final pure state possesses an exact Planckian distribution. As has been shown in Kiefer (2001), the coupling of this pure state (a squeezed state in quantum-optics language) to its natural environment produces a thermal ensemble for the Hawking radiation, which constitutes an open quantum system, after this environment has been traced out. The thermal nature of this radiation is thus a consequence of decoherence (Giulini *et al* 1996).

There exist many attempts to derive the Bekenstein–Hawking entropy (7.22) within approaches to quantum gravity, see, e.g., Kiefer (1999) and Wald (2001) for more details and references. Examples are the derivations within superstring theory (counting of states referring to microscopic objects called D-branes), canonical quantum gravity, Sakharov's induced gravity, conformal field theories,

and others. Although many of these look very promising, a final consensus has not yet been reached.

7.5 Primordial black holes

Can the previously discussed quantum effects of black holes be observed? As has already been mentioned, black holes formed by stellar collapse are much too heavy to exhibit quantum behaviour. To form smaller black holes one needs higher densities which can only occur under the extreme situations of the early Universe[3]. Such primordial black holes can originate in the radiation-dominated phase during which no stars or other objects can be formed.

Consider for simplicity a spherically symmetric region with radius R and density $\rho = \rho_c + \delta\rho$ embedded in a flat Universe with the critical density ρ_c, cf Carr (1985). For spherical symmetry the inner region is not affected by matter in the surrounding part of the Universe, so it will behave like a closed Friedmann Universe (since its density is overcritical), i.e. the expansion of this region will come to a halt at some stage, followed by a collapse. In order to reach a complete collapse, the (absolute value of the) potential energy, V, at the time of maximal expansion has to exceed the inner energy, U, given by the pressure p, that is,

$$V \sim \frac{GM^2}{R} \sim G\rho^2 R^5 \gtrsim pR^3. \tag{7.29}$$

If the equation of state reads as $p = wR$ ($w = 1/3$ for radiation dominance), this gives

$$R \gtrsim \sqrt{w}\frac{1}{\sqrt{G\rho}}. \tag{7.30}$$

The lower bound for R is thus just given by the *Jeans length*. An upper bound also exists. The reason is that R must be smaller than the curvature radius (given by $1/\sqrt{G\rho}$) of the overdense region at the moment of collapse. Otherwise the region would contain a compact three-sphere which is topologically disconnected from the rest of the Universe. This case would not then lead to a black hole within our Universe. Using $\rho \sim \rho_c \sim H^2/G$, where H denotes the Hubble parameter of the background flat Universe, one has the condition

$$H^{-1} \gtrsim R \gtrsim \sqrt{w}H^{-1} \tag{7.31}$$

evaluated at the time of collapse, for the formation of a black hole. This relation can also be rewritten as a condition referring to any initial time of interest (Carr 1985). In particular, one is often interested in the time where the fluctuation enters the horizon in the radiation-dominated Universe. This is illustrated in figure 7.3, where the presence of a possible inflationary phase at earlier times is also shown.

[3] In theories with large extra dimensions it is imaginable that the quantum effects of black holes can be seen at ordinary accelerators, see Dimopoulos and Landsberg (2001).

Figure 7.3. Time development of a physical scale $\lambda(t)$ and the Hubble horizon $H^{-1}(t)$. During an inflationary phase $H^{-1}(t)$ remains approximately constant. After the end of inflation (a_f) the horizon $H^{-1}(t)$ increases faster than any scale. Therefore λ_k enters the horizon again at $t_{k,\text{enter}}$ in the radiation- (or matter-) dominated phase.

At horizon entry one gets, denoting $\delta \equiv \delta\rho/\rho_c$,

$$1 \gtrsim \delta_{\text{enter}} \gtrsim 0.3. \tag{7.32}$$

This is, however, only a rough estimate. Numerical calculations give instead the bigger value of $\delta_{\min} \approx 0.7$ (Niemeyer and Jedamzik 1999).

Taking from (7.31) $R \approx \sqrt{w}H^{-1}$, one gets for the initial mass of a primordial black hole (PBH)

$$M_{\text{PBH}} = \frac{4\pi}{3}\rho R^3 \approx \frac{4\pi}{3}\rho_c(1+\delta)w^{3/2}H^{-3} \approx w^{3/2}M_H \tag{7.33}$$

where $M_H \equiv (4\pi/3)\rho_c H^{-3}$ denotes the mass inside the horizon. Since M_{PBH} is of the order of this horizon mass, a collapsing region will form a black hole practically immediately after horizon entry. Using the relation $M_H = t/G$, valid for a radiation-dominated Universe, one gets from (7.33) the quantitative estimate

$$M_{\text{PBH}} \, [\text{g}] \approx 10^{38} t \, [\text{s}]. \tag{7.34}$$

This means that one can create Planck-mass black holes at the Planck time, and PBHs with $M_{\text{PBH}} \approx 5 \times 10^{14}$ g at $t \approx 5 \times 10^{-24}$ s. The latter value is important since, according to (7.18), black holes with masses smaller than $M_{\text{PBH}} \approx 5 \times 10^{14}$ g have by now evaporated due to Hawking radiation. PBHs with bigger mass are still present today. At $t \approx 10^{-5}$ s, one can create a solar-mass black hole and at $t \approx 10$ s (the time of nucleosynthesis) one could form a

PBH with the mass of the Galactic black hole. The initial mass can increase by means of accretion, but it turns out that this is negligible under most circumstances (Carr 1985).

In the presence of an inflationary phase in the early Universe, all PBHs produced before the end of inflation are diluted away. This gives the bound

$$M_{\text{PBH}} > M_{\text{H}}(T_{\text{RH}}) \approx \frac{m_p^3}{10.88 T_{\text{RH}}^2} \sim 1 \text{ g} \qquad (7.35)$$

if for the reheating temperature T_{RH} a value of 10^{16} GeV is chosen.

According to the numerical calculations by Niemeyer and Jedamzik (1999), there exists a whole spectrum of initial masses,

$$M_{\text{PBH}} = K M_{\text{H}} (\delta - \delta_{\min})^{\gamma} \qquad (7.36)$$

a relation that is reminiscent of the theory of critical phenomena. This may change some of the quantitative conclusions.

To calculate the production rate of PBHs, one needs an initial spectrum of fluctuations. This is usually taken to be of a Gaussian form, as predicted by most inflationary models (cf Liddle and Lyth 2000). Therefore, there always exists a non-vanishing probability that the density contrast is high enough to form a black hole, even if the maximum of the Gaussian corresponds to a small value. One can then calculate the mass ratio (compared to the total mass) of regions which will develop into PBHs with mass $M_{\text{PBH}} \gtrsim M$, see, e.g., Bringmann *et al* (2001 section 2) for details. This mass ratio, given by

$$\alpha(M) := \frac{\rho_{\text{PBH},M}}{\rho_{\text{r}}} \approx \Omega_{\text{PBH},M} \equiv \frac{\rho_{\text{PBH},M}}{\rho_{\text{c}}} \qquad (7.37)$$

where ρ_{r} is the radiation density, is then compared with observation. This, in turn, gives a constraint on the theoretically calculated initial spectrum. Table 7.2 presents various observational constraints on α (see Green and Liddle 1997). The corresponding maximal value for each α is shown for the various constraints in figure 7.4.

Constraints arise either from Hawking radiation or from the gravitational contribution of PBHs to the present Universe (last entry). PBHs with initial mass of about 5×10^{14} g evaporate 'today'. (They release about 10^{30} erg in the last second.) From observations of the γ-ray background one can find the constraint given in table 7.2 which corresponds to an upper limit of about 10^4 PBHs per cubic parsec or $\Omega_{\text{PBH},0} < 10^{-8}$. One can also try to observe directly the final evaporation event of a single PBH. This gives an upper limit of about 4.4×10^5 events per cubic parsec per year.

Given these observational constraints, one can then calculate the ensuing constraints on the primordial spectrum. The gravitational constraint $\Omega_{\text{PBH},0} < 1$ gives surprisingly strong restrictions (cf Bringmann *et al* 2001). For a scale-free spectrum of the form $\propto k^n$, as is usually discussed for inflationary models,

Table 7.2. Constraints on the mas fraction $\alpha(M) := \frac{\rho_{\mathrm{PBH},M}}{\rho_{\mathrm{r}}} \approx \Omega_{\mathrm{PBH},M}$ of primordial black holes at their time of formation (Green and Liddle 1997).

Constraint	Range	Reason
$\alpha < 0.1(M/10^{15} \text{ g})^{\frac{3}{2}}$	$M < 10^{15}$ g	Radiation relics
$\alpha < 10^{-17}(10^9 \text{ g}/M)^{\frac{1}{2}}$	$10^9 \text{ g} < M < 10^{11}$ g	$n_{\mathrm{n}}/n_{\mathrm{p}}$-ratio
$\alpha < 10^{-22}(M/10^{10} \text{ g})^{\frac{1}{2}}$	$10^{10} \text{ g} < M < 10^{11}$ g	Deuterium dissociation
$\alpha < 10^{-21}(M/10^{11} \text{ g})^{\frac{5}{2}}$	$10^{11} \text{ g} < M < 10^{13}$ g	Helium fission
$\alpha < 10^{-16}(10^9 \text{ g}/M)$	$10^9 \text{ g} < M < 10^{13}$ g	Entropy per baryon
$\alpha < 10^{-26}$	$M \approx 5 \times 10^{14}$ g	γ background
$\alpha < 10^{-18}(M/10^{15} \text{ g})^{\frac{1}{2}}$	$M > 10^{15}$ g	Present PBH density

Figure 7.4. Strongest constraints on the initial PBH mass fraction. The numbers correspond to the various entries in table 7.2.

one finds restrictions on n that are comparable to the limits obtained by large-scale observations (the anisotropy spectrum of the cosmic microwave background radiation). Since these restrictions come from observational constraints referring to much smaller scales, they constitute an important complementary test.

The question as to whether PBHs really exist in nature has thus not yet been settled. Their presence would be of an importance that could hardly be overestimated. They would give the unique opportunity to study the quantum effects of black holes and could yield the crucial key for the construction of a final theory of quantum gravity.

References

Bekenstein J D 1980 *Physics Today* (January) 24

Bringmann T, Kiefer C and Polarski D 2001 *Phys. Rev.* D **65** 024008

Carr B J 1985 *Observational and Theoretical Aspects of Relativistic Astrophysics and Cosmology* ed J L Ganz and L J Goicoechea (Singapore: World Scientific) p 1

Chen P and Tajima T 1999 *Phys. Rev. Lett.* **83** 256

Davies P C W 1975 *J. Phys. A: Math. Gen.* **8** 609

Dimopoulos S and Landsberg G 2001 *Phys. Rev. Lett.* **87** 161602

Farrugia Ch J and Hajicek P 1979 *Commun. Math. Phys.* **68** 291

Frolov V P and Novikov I D 1998 *Black Hole Physics* (Dordrecht: Kluwer)

Fulling S A 1973 *Phys. Rev.* D **7** 2850

Giulini D, Joos E, Kiefer C, Kupsch J, Stamatescu I O and Zeh H D 1996 *Decoherence and the Appearance of a Classical World in Quantum Theory* (Berlin: Springer)

Green A M and Liddle A R 1997 *Phys. Rev.* D **56** 6166

Hawking S W 1975 *Commun. Math. Phys.* **43** 199

Hehl F W, Kiefer C and Metzler R (eds) 1998 *Black Holes: Theory and Observation* (Berlin: Springer)

Heusler M 1996 *Black Hole Uniqueness Theorems* (Cambridge: Cambridge University Press)

Kiefer C 1994 *Canonical Gravity: From Classical to Quantum* ed J Ehlers and H Friedrich (Berlin: Springer) p 170

——1999 *Classical and Quantum Black Holes* ed P Fré *et al* (Bristol: Institute of Physics Publishing) p 18

——2001 *Class. Quantum Grav.* **18** L151

Leinaas J M 2001 hep-th/0101054

Liddle A R and Lyth D H 2000 *Cosmological Inflation and Large-Scale Structure* (Cambridge: Cambridge University Press)

Niemeyer J C and Jedamzik K 1999 *Phys. Rev.* D **59** 124013

Núñez D, Quevedo H and Sudarsky D 1998. See Hehl *et al* 1998 p 187

Page D N 1976 *Phys. Rev.* D **13** 198

Parikh M K and Wilczek F 2000 *Phys. Rev. Lett.* **85** 5042

Sexl R U and Urbantke H K 1983 *Gravitation und Kosmologie* (Mannheim: Bibliographisches Institut)

Thorne K S 1994 *Black Holes and Time Warps* (New York: Norton)

Unruh W G 1976 *Phys. Rev.* D **14** 870

Wald R M 1984 *General Relativity* (Chicago, IL: University of Chicago Press)

——2001 *The Thermodynamics of Black Holes* http://www.livingreviews.org/Articles/Volume4/2001-6wald

This page intentionally left blank

PART 3

OUR GALACTIC CENTER

This page intentionally left blank

Chapter 8

The mass of the Galactic Center black hole

Andreas Eckart
Universität zu Köln, Germany

With our knowledge of black holes and the Galactic structure from chapters 1, 2, and 4, we can now address the Galactic Center black hole itself. The investigation of stellar velocities close to the Galactic Center has become the primary means for determining the mass and mass density of this black hole. This chapter recalls the history of near-infrared observations of the Galactic Center, describes the near-infrared speckle imaging technique used recently, and presents the results and prospects of imaging and spectroscopy of stars near the black hole.

8.1 Introduction and summary

At a distance of only 8 kpc the Galactic Center is the closest nucleus of a galaxy, 100 to 1000 times closer than the nearest extragalactic systems. It is thus a unique laboratory in which physical processes that are also relevant for nuclei of other galaxies can be studied with the highest angular resolution possible. The gas and dust in the line of sight to the center, however, make it impossible to observe it at optical or UV wavelengths. It is only accessible in the radio, infrared and X-ray domain. The optical extinction amounts to 30 magnitudes ($A_V = 30^{mag}$) or an attenuation of visible light by a factor of 10^{-12}. This corresponds in the near-infrared K band at a wavelength of 2.2 μm to an extinction of only $A_K = 3.3^{mag}$.

Subarcsecond-angular-resolution line imaging (Eckart *et al* 1995, Tamblyn *et al* 1996) and 3D imaging spectroscopy (Krabbe *et al* 1995, Genzel *et al* 1996) have shown that several of the brightest members of the central cluster—the so called IRS16 complex—are He I emission line stars. These objects show in their spectra prominent broad He I and H I recombination lines in emission. From non-LTE (LTE = local thermal equilibrium), stellar atmosphere modeling of the observed emission characteristics of several of the He I stars Najarro *et al* (1994, 1997) have inferred that these objects are moderately hot (17 000–30 000 K), very

229

luminous (1–30×10^5 L$_\odot$) massive stars whose helium-rich surface layers expand as powerful stellar winds with velocities of 200–800 km s^{-1} and mass-loss rates of 1–70×10^{-5} M$_\odot$/yr. These stars have formed within the last few million years and provide the dominant fraction of the total luminosity of the central parsec. Imaging spectroscopy shows that within the central 0.3–0.4 pc of the stellar cluster bright late-type stars (supergiants and the brightest AGB stars) are absent but that the core is surrounded by a ring of red supergiants/AGB stars showing strong CO-band head absorptions in their near-infrared spectra (Genzel *et al* 1996, see also earlier work by Sellgren *et al* 1990, Haller *et al* 1996).

The gas and stellar velocities increase toward the position of Sagittarius A* (Sgr A*: see chapter 11) indicating the presence of a large compact mass. The evidence for a dark central mass of 1–3×10^6 M$_\odot$ in the Galactic Center has been steadily growing over the past two decades from observations of radial velocities of gas and stars (Lacy *et al* 1980, Serabyn and Lacy 1985, Genzel *et al* 1985, Sellgren *et al* 1990, Krabbe *et al* 1995, Haller *et al* 1996, Genzel *et al* 1996). A reliable estimate of the enclosed mass and its compactness, however, can only be made if the full velocity field (radial and transverse components) is known (e.g. Genzel *et al* 2000, Ghez *et al* 2000). Furthermore, stars as tracers of the velocity field are much more reliable than gas, since they are not influenced by magnetic pressure or frictional forces. Using stars as tracers of the gravitational potential therefore allows us to measure the mass content and mass concentration within the central parsec of our Galaxy. Recent review articles on the phenomena and physical properties of the Galactic Center are: Genzel and Townes (1987), Genzel *et al* (1994), Blitz *et al* (1993), Genzel (1989), Mezger *et al* (1996), Morris and Serabyn (1996), and Melia and Falcke (2001).

This chapter summarizes results of a program to determine stellar velocities in the plane of the sky (proper motions). From high-resolution near-infrared imaging over the past 10 years proper motions for about 100 stars between about 0.01 and 0.3 pc from the compact radio source Sgr A* have been determined. The proper motion and radial velocity dispersions are in very good agreement indicating that the stellar velocity field is, on average, close to isotropic. Comparing individual images from different observing epochs one finds within the central arcsecond (0.04 pc) several fast moving stars with velocities in excess of 1000 km s^{-1} in the immediate vicinity (0.01 pc) of Sgr A*. From the stellar radial and proper motion data, one can infer that a dark mass of $2.9 \pm 0.35 \times 10^6$ M$_\odot$ must reside within about a light week of the compact radio source. Its density must be about 4×10^{12} M$_\odot$ pc^{-3} or greater (Genzel *et al* 2000).

A direct link between the radio and infrared positional reference frames via the maser emission of five infrared bright stars leads to the first possible detection of a near-infrared source at the position of Sgr A*. Speckle spectroscopy observations of sources within the central arcsecond allow an estimate of their stellar masses and via an equipartition argument a determination of a lower limit to the mass of about 10^{3-5} M$_\odot$ that are associated with Sgr A* itself (Reid *et al* 1999). There is no stable configuration of normal stars, stellar remnants

(neutron stars or stellar black holes) or sub-stellar entities that can account for the combination of both—the high mass and the high mass density (Maoz 1998). The combination of these data—especially the proper motion measurements of the stars in the central arcsecond—now provides compelling evidence for a massive black hole at the core of the Milky Way. In order to put the presented data in a broader context I summarize in the following the history of imaging the Galactic Center in the near-infrared.

8.2 A brief history of imaging the Galactic Center in the near-infrared

Attempts to detect a source at the position of the Galactic Center in the near infrared started as early as 1945 (Stebbins and Whitford 1947, Moroz 1961). Due to a combination of a lack in sensitivity and coarse sampling these initial efforts were not successful. The first detection was achieved by Becklin and Neugebauer (1968) at a wavelength of 2.2 μm in scans with 0.25$'$ and 0.08$'$ apertures corresponding to linear resolutions of 0.62 and 0.2 pc. These scans revealed the compact nuclear stellar cluster for the first time. In the following years single detector maps with higher spatial resolution were obtained by a number of authors. These maps resolved the central cluster into individual bright complexes. The introduction of multiplexed, near-infrared array detectors allowed more efficient mapping with yet higher angular resolution. The first maps of the Galactic Center using array detectors were obtained by Forrest *et al* (1986). These measurements started to resolve the central IRS16 complex into many individual sources. Lunar occultation measurements (Simons *et al* 1990, Simon *et al* 1990, Adams *et al* 1988) demonstrated that the brightest sources in the IRS 16 complex (IRS 16NE, 16C, 16SW and 16NW) are very compact with diameters less than 100 AU and therefore are most likely individual or multiple stars but not large clusters. Two-dimensional speckle imaging (as presented here and first published in Eckart *et al* 1992 and subsequent papers) resulted in the first diffraction-limited maps (0.15$''$ angular resolution corresponding to 0.006 pc) of the central 20$''$ × 20$''$ at 2.2 μm and 1.6 μm. So far these measurements have resolved the central cluster including several compact stellar complexes like IRS 1, IRS 13, and IRS 16SW complexes into about 600 individual stars (Eckart *et al* 1995). They have also revealed a complex of near-infrared sources very close to the position of Sgr A* for the first time (Genzel *et al* 1997). These results have been confirmed by repeated speckle imaging at the NTT, observations under excellent seeing conditions (Herbst *et al* 1993) as well as the first 'tip-tilt' (a crude adaptive optics technique to compensate for seeing) measurements (Close *et al* 1995) at 1.6 μm with a resolution of 0.3$''$. These latter observations, however, have not been able to resolve the very central stellar cluster. Only recent speckle imaging with the 10 m diameter Keck telescope (Klein *et al* 1996, Ghez *et al* 1998) as well as the first adaptive optics measurements (Gezari *et al*

2000 and, most recently, the results from the ESO VLT adaptive optics: NAOS CONICA, in preparation) have revealed the same subarcsecond structure for the small stellar cluster around the radio position of Sgr A*. These observations have also confirmed the velocities of the stars in this area. With these techniques available it will be possible in the next years to exploit the Galactic Center as a laboratory for studying the physical processes in the immediate vicinity of a massive black hole. Future measurement with 8–10 m class telescopes as well as near-infrared interferometric measurements will result in a determination of the full three-dimensional orbits of the stars in the center. They will also help to monitor the variable source at the position of Sgr A*, will allow us to further constrain physical models of black hole accretion, and help to search for potential lensing effects of background stars by the central black hole (see Alexander and Sternberg 1999, Alexander and Loeb 2001; and chapter 9).

8.3   Speckle interferometry

Detailed descriptions of (infrared) speckle interferometry and imaging have been given in several reviews (Christou 1991, Roddier 1989, Mariotti 1989, Dainty 1975). Here I only give a brief summary in order to introduce expressions that are linked to this technique and that are used in the following description of the data reduction and interpretation. The atmospheric turbulence above the telescope distorts the otherwise plane stellar wavefronts and is responsible for a point spread function that varies rapidly with time. This phenomenon is called seeing. Over entities of the turbulent layer—seeing cells with a Fried (1966) diameter $r_0 \sim 20$–50 cm in the near-infrared (NIR)—the phase front predominantly experiences a linear phase change only. Neighbouring stars within the isoplanatic patch of about a 20–30$''$ diameter in the NIR have a similar resulting PSF (point spread function) within the coherence time of the atmosphere (up to a few 100 ms in the NIR).

An image $I(x, y)$ of an object $O(x, y)$ taken in the focal plane (with coordinates x and y) through the combination of the telescope and the turbulent atmosphere can be written as the convolution of $O(x, y)$ with the combined telescope and atmospheric PSF $P(x, y)$:

$$I(x, y) = O(x, y) * P(x, y). \tag{8.1}$$

Co-adding all short exposures after shifting the positions of the brightest speckle in the seeing cloud $I(x, y)$ of a bright reference object to a common location results in an image containing substantial power at the diffraction limit of the telescope. The speckle image processing can formally be described in the following way: two-dimensional images $I(x, y)$ can be written as the object $O(x, y)$ convolved with the combined telescope and atmospheric PSF $P(x, y)$ plus an additive noise component $N(x, y)$,

$$I(x, y) = O(x, y) * P(x, y) + N(x, y). \tag{8.2}$$

In the simple shift-and-add algorithm one determines the location of the brightest speckle in the seeing disk of a reference star in each of the M images $I_m(x, y)$ of a series of short exposures and shifts them to the same position before summing the images:

$$S(x, y) = \frac{1}{M} \sum_{m=1}^{M} I_m(x + x_m, y + y_m).$$ (8.3)

The x_m and y_m are the corresponding shift vector components. One obtains a shift-and-add image $S(x, y)$ that can be written as the object convolved with a shift-and-add PSF $P_s(x, y)$ plus a modified noise contribution $N_s(x, y)$:

$$S(x, y) = \frac{1}{M} \left(O(x, y) * \sum_{m=1}^{M} P_m(x + x_m, y + y_m) + \sum_{m=1}^{M} N_m(x + x_m, y + y_m) \right)$$ (8.4)

or

$$S(x, y) = \frac{1}{M} (O(x, y) * P_s(x, y) + N_s(x, y)).$$ (8.5)

For each star in the field of view (smaller than the isoplanatic patch size) this PSF results in a diffraction-limited image of the object on top of a residual seeing background built up by the contributions from all fainter speckles. In contrast to the long exposure (a simple co-addition of individual frames) the raw shift-and-add image contains information at the diffraction limit of the telescope. In the NIR and especially for the Galactic Center observations the shift-and-add algorithm has the following advantages over the other well-known speckle reduction algorithms which are based on a Fourier analysis of the images like the Knox–Thompson method and bi-spectrum analysis: (1) no Fourier transforms have to be involved resulting in a fast processing of a large amount of data; and (2) local artifacts in the image plane are not spread all over the Fourier plane making it easier to employ deconvolution methods especially if the bright supergiant IRS 7 is in the same field of view as the Sgr A* region.

After application of the shift-and-add algorithm the resulting images contain up to approximately 20% of the image power at the diffraction limit but they still have to be corrected for the PSF produced by the shift-and-add algorithm. In the data reduction scheme presented here this is done using the Lucy algorithm (Lucy 1974). This correction aims at replacing the PSF of the shift-and-add algorithm that contains a broad seeing pedestal in addition to a diffraction-limited core with a single Gaussian-shaped PSF.

8.4   The center of the Milky Way

8.4.1   Imaging and proper motions

Over the past 10 years or so we have been conducting a programme to study the properties of the central nuclear stellar cluster via NIR high spatial resolution
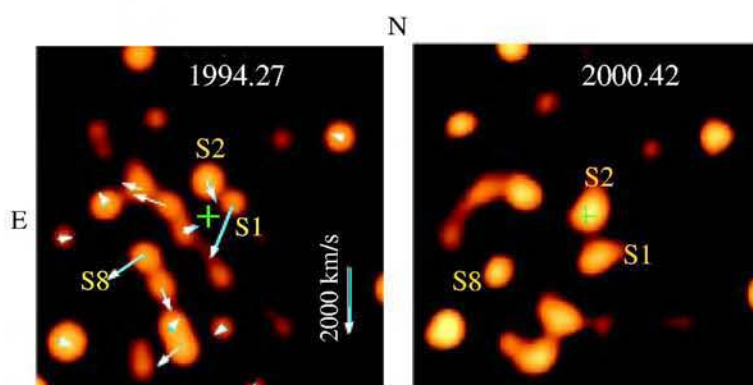
Figure 8.1. A comparison of two diffraction-limited images taken at different epochs—1994.27 and 2000.42—using the MPE SHARP camera at the ESO NTT. The cross marks the position of the radio source Sgr A*. The arrows indicate the velocities of stars in the central $1.5'' \times 1.5''$. Sources S1, S2, and S8 are labeled. See also color section.

measurements using the MPE (Max-Planck-Institut für extraterrestrische Physik) speckle camera SHARP at the 3.5 m New Technology Telescope (NTT) of the European Southern Observatory (ESO). This program has resulted in the very first detection of the proper motions of stars that correspond to velocities of up to 1400 km s^{-1} in the central arcsecond in the vicinity of Sgr A* (Eckart and Genzel 1996, 1997). These results have been confirmed by Ghez *et al* (1998). On the 1.5–3σ level we have now detected (Eckart *et al* 2001, 2002) orbital curvatures which confirm the recent results by Ghez *et al* (2000).

In figure 8.1 we show the stars in the central $2''$ for two representative epochs: 1994.27 and 2000.42. The position of the compact radio source Sgr A* is indicated by a central cross. In the image on the left-hand side the velocities are shown as vectors with their end points at the position of the corresponding stars at the later epoch. A comparison of both images clearly shows that the density of sources especially in the central arcsecond is high enough that at the currently reached point source sensitivity (mostly limited by the wings of bright neighboring stars) this region needs to be monitored at least once per year. Such a dense monitoring allows us to identify and track the sources without any doubt. For the fastest star—S1—the angular velocity of almost 40 mas yr^{-1} corresponds to a linear velocity of 1400 km s^{-1}.

The analysis of our best images over the past 10 years has resulted in the detection of the proper motions of stars within the central arcsecond. These motions translate into a one-dimensional velocity dispersion of the stars in the central arcsecond (corrected for the measurement error) of the order of >400 km s^{-1} (within a radius of $\sim 0.5''$). Of the 2×104 proper motions in Genzel *et al* (2000), 48 (23%) are determined to 4σ or better. Five proper motions are determined at the $\sim 10\sigma$ level. Of the 227 line-of-sight velocities, 38 (17%) are

determined to 4σ or better. For at least 14 (of 29) He I emission-line stars and for 18 late-type stars we now have determinations of all three velocity components. With the exception of small amounts of anisotropy (Genzel *et al* 2000; which are taken into account in Jeans modeling; see later) between 0.035 and 0.35 pc from the compact radio source Sgr A* the projected proper motion velocity dispersion per coordinate is in excellent agreement with the radial velocity dispersion results.

Overall the stellar motions do not deviate strongly from isotropy and are consistent with a spherical isothermal stellar cluster. However, a small deviation from isotropy is found for the sky-projected velocity components of the young, early-type stars. Most of the bright He I emission line stars are on apparently tangential orbits. This overall rotation could be a remnant of the original angular momentum pattern in the interstellar cloud from which these stars were formed. The fainter, fast moving stars within $\sim 1''$ from Sgr A* (the 'Sgr A* cluster') appear to be moving largely on radial orbits. Speckle spectroscopy with SHARP at the NTT (Genzel *et al* 1997) and slit spectroscopy with ISAAC at the VLT suggest that several of them are early-type stars. This is consistent with the idea that these stars are members of the early-type cluster with small angular momentum and therefore fell into the immediate vicinity of Sgr A* (Genzel *et al* 2000, Gerhard 2000).

8.4.2 Spectroscopy

Eckart *et al* (1999) reported results based on new NIR observations of the central stellar cluster of our Galaxy conducted with the infrared spectrometer ISAAC at the ESO VLT UT1 and the MPE speckle camera SHARP at the ESO NTT (see also Herbst *et al* 1993, Figer *et al* 2000). The ISAAC observations resulted in $\lambda/\Delta\lambda \sim 5000$ K-band spectra of the 2.058 μm He I, 2.165 μm Br γ emission lines, and 2.29 μm CO-bandhead absorption line (see figure 8.2). These data clearly demonstrate that there is no strong CO-bandhead absorption originating in the northern part (S1/S2 area) of the central stellar cluster at the position of Sgr A*. This makes it likely that these K ~ 14.5 stars are O9–B0.5 stars with masses of 15 to 20 $M_\odot$. Weaker CO-bandhead absorption in the southern part of the cluster (S10/S11 area) is most likely due to contributions from neighbouring stars. Eckart *et al* (1999) also report the detection of Br γ line emission at the position of the central stellar cluster which could be associated with the 'mini-spiral' rather than with the Sgr A* cluster itself.

8.4.3 Enclosed mass

Together with the VLBI maser nucleus of NGC 4258 (Greenhill *et al* 1995, Myoshi *et al* 1995) the compact dark mass in the Galactic Center is currently the best and most promising case for a supermassive nuclear black hole (Maoz 1998). The new anisotropy-independent mass estimates (Leonard–Merritt estimators of the proper motions) as well as Jeans modeling (explicitly including velocity
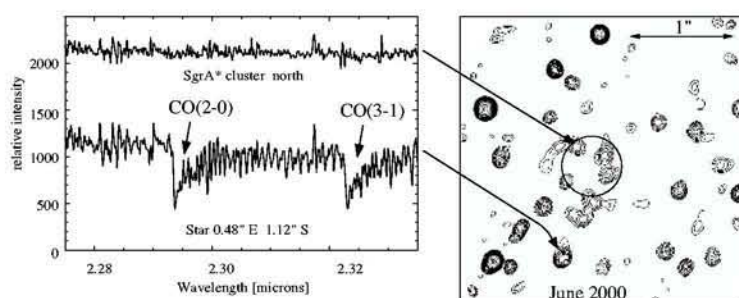
Figure 8.2. Spectra of the CO(2–0) and CO(3–1) band head absorption toward the northern part of the central Sgr A* stellar cluster and a star to the south of it. The star and the northern area are indicated in a contour plot of the 2 μm continuum emission map from the central $2.9'' \times 2.9''$ as obtained with SHARP at the NTT.

anisotropy, see figure 8.3) result in a compact (<0.0058 pc) mass close to 2.9×10^6 M$_\odot$ with a mass density greater than 4×10^{12} M$_\odot$ pc^{-3} (Genzel *et al* 2000, see also Eckart *et al* 2001, 2002). One can show that any cluster of that mass at such a high density cannot be stable over more than 10^6–10^7 years (Maoz 1998). Equipartition arguments that include the known proper motions of the radio source Sgr A* (<16 km s^{-1}, Backer 1996, Reid *et al* 1999, Genzel *et al* 2000) and the estimated mass and known proper motion of the inner fast moving stars (Eckart and Genzel 1997, Genzel *et al* 1997) result in a lower limit of at least 10^3 M$_\odot$ that has to be associated with Sgr A*. The current conclusion is that this mass is most likely contained in a single massive black hole.

Due to the limited number of detected stars we currently use a minimum radius for our determination of the mass and mass density of 0.01 pc (0.25″). The $\alpha = 5$ Plummer (see later) model of a dark cluster results in a core radius of such a hypothetical cluster of $r_0 = 0.0058$ pc (0.23″) and corresponding central density of >4 $\times 10^{12}$ M$_\odot$ pc^{-3} (see earlier). The star S2 is currently at a distance of less than 0.1 from the center—four times closer than the minimum radius previously mentioned. If the orbit of S2 remains consistent with a central unresolved mass of 2.9×10^6 M$_\odot$ the mass density is at least 64 times higher, i.e. 2.4×10^{14} M$_\odot$ pc^{-3}. In this case the collapse life time would shrink to only a few 10^6 years, making the Galactic Center the strongest of all massive black hole candidates.

Figure 8.3 shows the mass distribution in the central 10 pc of the Galaxy obtained from stellar and gas dynamics (for $R = 8.0$ kpc). Bold 'G's denote mass estimates from ionized and neutral gas dynamics. Rectangles with crosses and downward-pointing triangles denote the isotropic mass modeling of Genzel *et al* (1996, 1997), including Jeans modeling of stellar radial velocities (early- and late-type stars, filled downward-pointing triangles) and Bahcall–Tremaine estimators of the NTT proper motions until 1996 (open down-pointing triangles; Eckart and Genzel 1997). Open rectangles (with crosses) are Bahcall–Tremaine estimators
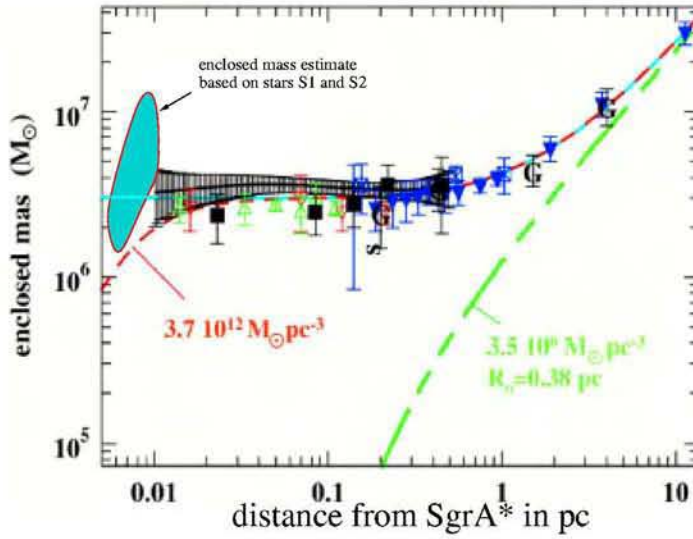
Figure 8.3. The enclosed mass contained in circular apertures plotted as a function of their radius. The apertures are centered on the position of Sgr A*. An explanation is given in the text. See also colour section.

of the line-of-sight-velocity data only. Open upward-pointing triangles are the Bahcall–Tremaine estimators of the 1995–96 proper motion data of Ghez *et al* (1998). The new anisotropy-independent mass estimates from the present work are given as filled black rectangles from Leonard–Merritt estimators (Leonard and Merritt 1989) of the proper motions and as large black crosses connected by a continuous curve resulting from the Jeans modeling. For comparison several model curves are shown. The dashes represent the mass model for the (visible) stellar cluster $M_L(2\,\mu m) = 2$, $r_{core} = 0.38$ pc, $\rho(r = 0) = 3.5 \times 10^6\,M_\odot\,pc^{-3}$. The continuous curve is the sum of this stellar cluster, plus a point mass of $2.9 \times 10^6\,M_\odot$. The short dashes are the sum of the visible stellar cluster, plus an $\alpha = 5$ Plummer model, $\rho(r) = \rho(0)[1 + (r/r_0)^2]^{-\alpha/2}$, of a dark cluster of central density $\sim 4 \times 10^{12}\,M_\odot\,pc^{-3}$ and $r_0 = 0.0058$ pc.

8.4.4   Orbital curvatures

For three sources S1, S2 and S8 we have detected a curvature in the orbits on the 1.5σ to 3σ level. In figure 8.4 we show the offset positions for S2 in declination and right ascension as a function of time. The linear fits to the first and second half of the data set clearly result in different slopes. Their difference divided by the time difference between the two intervals is a direct measure of the orbital curvature. In order to improve the statistics on an estimate of the accelerations we have used velocities derived from data covering different time intervals (Eckart *et*
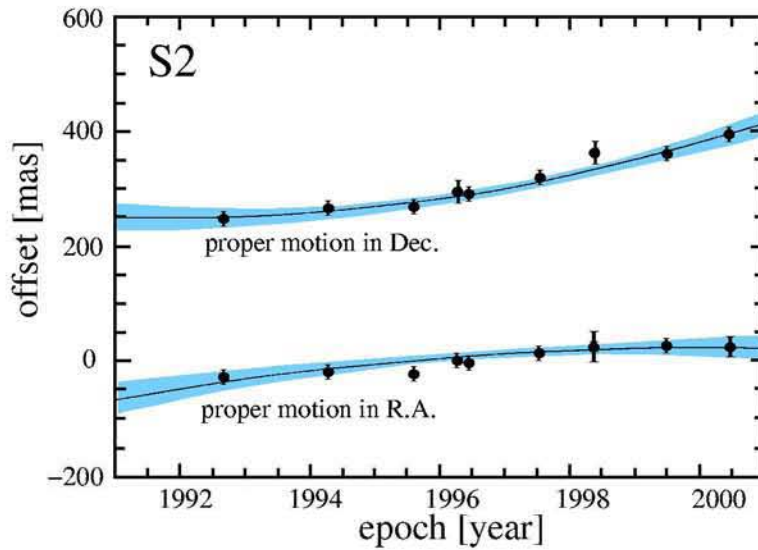
Figure 8.4. The relative positions of S2 in declination and right ascension as a function of the observing epoch. The declination velocity plot has been shifted by $+300$ km s^{-1}.

al 2001, 2002). For S2 we find a curvature of 2.3 ± 0.9 mas yr^{-2} corresponding to an acceleration of 95 km s^{-1} yr^{-1}. For S1 and S8 we find a curvature of 3.8 ± 2.4 mas yr^{-2} and 3.3 ± 1.1 mas yr^{-2}, respectively. The slopes are in agreement with those expected from fits of Keplerian orbits to the data (for S2 see figures 8.4 and 8.5). As an example the SHARP data now start to constrain the possible orbits for S2. For S2 a likely solution (obtained from least-square fits of Keplerian orbits to the data) is that this star has a line-of-sight offset of 8–9 mpc and a line-of-sight velocity in the range of -200 to -600 km s^{-1}. Due to current uncertainties in the inclination a combination of -8 to -9 mpc and 200–600 km s^{-1} is possible as well. Three possible orbits that represent good fits to the data are shown in figure 8.5. The current data and analysis indicate that S2 is aproaching its periastron.

The accelerations from the SHARP experiment (Eckart *et al* 2001, 2002) are consistent with recent results by Ghez *et al* (2000) and imply that the three stars orbit a central, compact mass (see figure 8.6).

Star S8 was excluded from the analysis, since the current proper motion velocity and radial separation from the center indicate that the measured acceleration requires orbital motion around a compact object with a mass in excess of 3×10^6 M$_\odot$. The data suggest that *either* this star was or is subject to a close interaction with a different object *or* that its position measurements are influenced by the emission of a different cluster star. Therefore, the analysis of the enclosed mass is solely based on the available data for S1 and S2. For each of
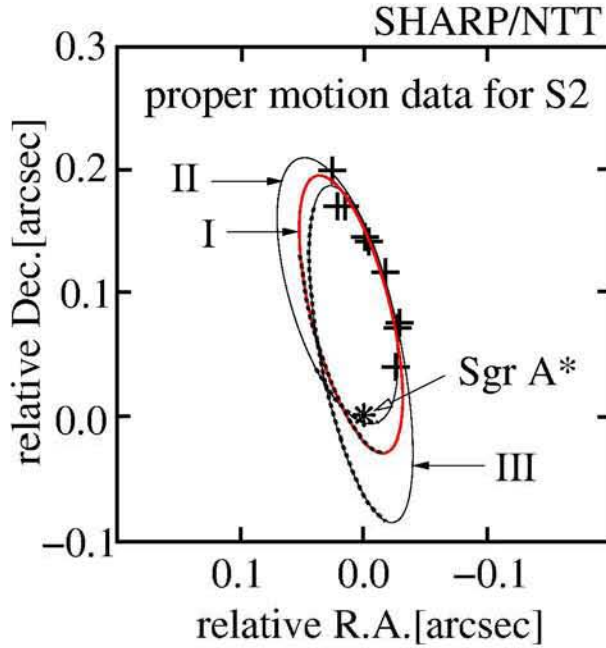
Figure 8.5. Three inclined Keplerian orbits that have line-of-sight velocities and separations from Sgr A* as mentioned in the text. The orbits represent the range of best fits to the 1992 to 2000 SHARP data for the fast moving source S2. From the proper motions alone we find a velocity of greater than 860 km s^{-1}—the results of the orbit modeling suggest a full space velocity of up to ∼1050 km s^{-1}.

the two sources S1 and S2 the acceleration values define an acceleration vector at an angle ϕ that should point towards the central source. Here we assume that the probability for the location of the central mass is uniform in ϕ. In figure 8.6 the stars S1 and S2 have been plotted at their time-averaged position resulting from the corresponding data sets. The measurement uncertainties define an error cone. For the presentation in figure 8.6 the dashed and dotted lines indicate an error cone that corresponds to a width of 2σ in deviation from the nominal direction indicated by the acceleration vector. In order to determine the location of the central mass we perform a maximum likelihood (ML) analysis. As an ML score we use $\log(\mathrm{ML}) = -\chi_{S1}^2/2 - \chi_{S2}^2/2$. The thin contour lines in figure 8.6 indicate the locations at which $\log(\mathrm{ML})$ drops by 0.5 below the corresponding peak values. Here $\chi^2 = (\phi - \phi_0)^2/(\Delta\phi)^2$, ϕ_0 denotes the angle of the acceleration vector, ϕ the angle of any radial line within an error cone, and $\Delta\phi = \sigma$ the half-width of the cone. The central filled circle in figure 8.6 marks the radio position of Sgr A* and corresponding uncertainties of ±30 mas. Using the observed curvature value and the enclosed mass range of 2.6–3.3 × 10^6 M$_\odot$ imposes a limit on the projected
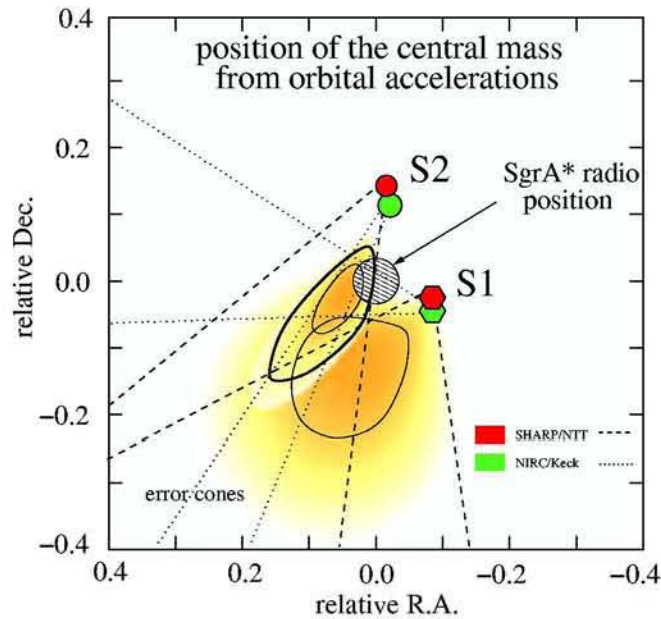
Figure 8.6. The acceleration vectors of the sources S1, S2 as derived from the current SHARP data. An explanation of the symbols is given in the text. The motion of the stars is consistent with orbits around a central $3 \times 10^6 M_\odot$ object. See also color section.

distance of S1 and S2 from Sgr A*. For the SHARP/NTT data this leads to an improvement in the determination of the Sgr A* position. We account for this effect in figure 8.6 by multiplying the log(ML) scores of the error cones with a Gaussian prior of the appropriate $1/e$ width centered on the time-averaged positions of S1 and S2. From the projection of the 1σ contour line (thin contour line east of the center in figure 8.6) the multiplied probabilities derived from the SHARP/NTT and NIRC/Keck data result in a position of a central dark mass of 48^{+54}_{-24} mas E and 18^{+42}_{-61} mas S of the nominal radio position of Sgr A*. Within these limits the central mass is located at the 68.5% confidence level ($\Delta\chi = 1.0$). At the 90% confidence level ($\Delta\chi = 2.71$) the central mass is located in an interval given by 48^{+109}_{-48} mas E and 18^{+72}_{-133} mas S (thick contour line left of the center in figure 8.6). Figure 8.6 shows that at the current ± 30 mas uncertainty of the radio position of Sgr A* the presently available accelerations of stars S1 and S2 alone are fully consistent with the hypothesis that the radio source Sgr A* is coincident with the center of the dark mass.

8.4.5 Is there an infrared counterpart of Sgr A*?

Menten *et al* (1996) have identified five H_2O/SiO maser stars within the central $20''$ of Sgr A* and the position of the radio source Sgr A* to within about
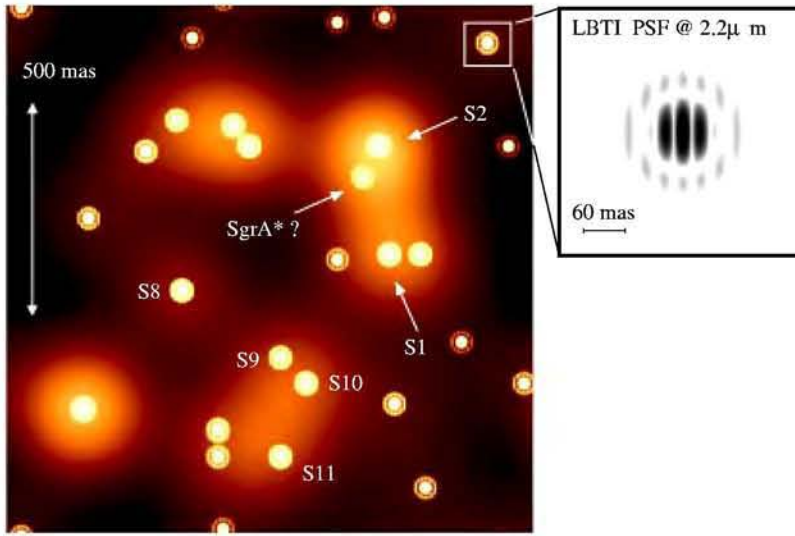
Figure 8.7. A simulation of how the central $1'' \times 1''$ of the Milky Way would be seen in 2000 with the LBT interferometer at a wavelength of 2.2 μm. Some of the fast moving stars are labeled. Given the large instantaneous field of view over which the high LBTI resolving power can be achieved the motions of the inner fast moving stars ($\sigma_{1 \text{ dim.}} \sim 400 \text{ km s}^{-1}$) can be referenced precisely to a large number (a few hundred objects with $m_K < 14$) of much slower ($\sigma_{1 \text{ dim.}} < 100 \text{ km s}^{-1}$) late-type stars in the central stellar cluster. A close-up view of the LBTI PSF is show in the top right-hand image. See also color section.

± 30 mas. It is located within the central stellar cusplike cluster. In 1994, 1995 and March/April 1996 it did not coincide with any of the $\approx 15^{\text{mag}}$ sources therein. However, in the SHARP June 1996 and the July 1997 data there is evidence for an additional source between S1, S2, and S3—right at the radio position of Sgr A*. In the diffraction-limited SHARP maps taken with a 50 mas and a 25 mas pixel scale this source manifests itself as an extension of S1 toward S2 (Genzel *et al* 1997). In the high SNR (resulting from a combination of several tens of thousands of frames) ~ 70 mas FWHM maps presented in Genzel *et al* (1997) the additional source is clearly separated from S1, S2, and S3. This object may represent the best candidate for the long-sought NIR Sgr A* counterpart.

In October 2000 the X-ray source at the position of Sgr A* underwent a large-amplitude X-ray flare and a weaker flare in September 1999 (Baganoff *et al* 2001a, b). In addition a cm-radio periodicity of 106 days was established from VLA data covering the past 20 years (Zhao *et al* 2001). Theoretical models (Markoff *et al* 2001) predict a flare K-magnitude between 8 and 13—about 0.5– 7 magnitudes brighter than the stars in the $1''$ diameter Sgr A* cluster. Such a burst could explain the possible NIR counterpart that was reported by Genzel *et al* (1997). A new extended Chandra exposure was be carried out in May 2002
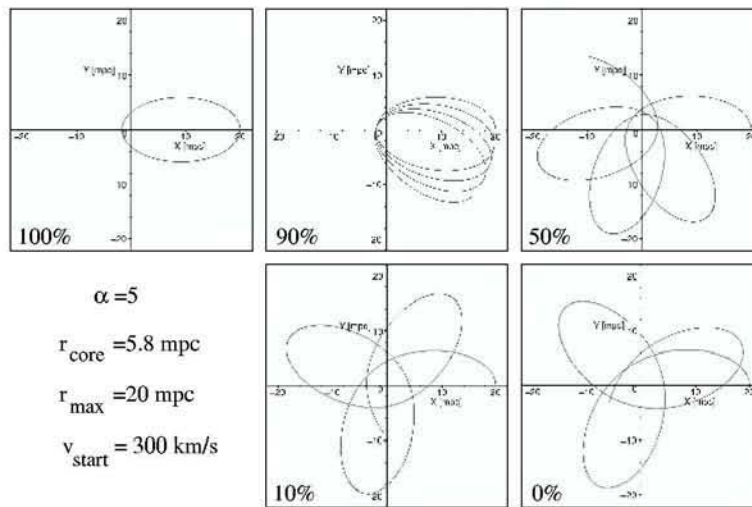
Figure 8.8. Possible stellar orbits if only the fraction of mass given in the panels is contained in a central unresolved object. The fraction not contained in such an object is here assumed to have a core radius of 5.8 mpc and a Plummer mass distribution with an exponent $\alpha = 5$ (see explanation in text and in the caption of figure 8.3). The stars were launched perpendicularly to the radius vector in the plane of the sky with a velocity of 300 km s^{-1} at a distance of 20 mpc from the center (Rubilar and Eckart 2000).

showing more flares. Simultaneous observations at radio to infrared wavelengths will help to teach us more about the possible emission mechanisms associated with bursts and quiescent emission of Sgr A*.

8.4.6 LBT and the Galactic Center

For experiments in the very near future the LBT (Large Binocular Telescope on Mount Graham in Arizona) is ideally suited to search for and monitor the flux of a Sgr A* NIR counterpart. High sensitivity combined with adaptive optics and interferometry using the two LBT mirrors will provide the required accuracy to separate the Sgr A* NIR counterpart from neighboring stars (figure 8.7). The unprecedented combination of high sensitivity and high angular resolution over a large field of view will allow significant motions in most of the 600 stars brighter than $m_K < 14$ covering the inner parsec of the central stellar cluster to be detected. Combined with imaging spectroscopy this may result in a large number of sources with measurements of all three velocity components. Full space velocities are essential to improve the analysis of the dynamical properties of the late-type stars and the inner bright higher velocity He I stars. This will undoubtedly help to determine the origin of the He I stars which may represent the remains of a dissolved young cluster (see Gerhard 2000). The spectra from fast moving stars

will be of special importance. Knowing their full space velocity will result in complete information on their orbits.

It is even more desirable to find and track the motion of stars that are as close to the center as possible. Orbital timescales at the resolution limit of the LBT interferometer could be in the range of a few months. A detection of a relativistic or Newtonian periastron shift would ultimately result in a determination of the compactness of the enclosed central mass (figure 8,8; Rubilar and Eckart 2000, see also Eckart *et al* 2001, 2002). The prograde relativistic periastron rotation is of the order of 17 arcminutes per revolution for a 60 mas (2.4 mpc; orbital timescale 6.8 yr) orbit with an eccentricity of $\epsilon = 0.9$. For a 15 mas (0.6 mpc; orbital timescale 0.9 yr) orbit with the same eccentricity the shift is already of the order of 1.1 degrees per revolution. Periastron shifts of 2 degrees could be observed with the LBTI with $\sim 1\sigma$ yr^{-1}.

When a small amount of the compact mass is extended the retrograde Newtonian periastron shift would be much larger. For orbits with a half axis as before and a modest eccentricity of $\epsilon = 0.5$ the shift may amount to several tens of degrees per revolution. This assumes that the extended mass is contained in a compact cluster of less than 6 mpc core radius. Comparing the relative magnitudes of the relativistic and Newtonian periastron shift one finds that if for S2-like orbits only about 0.1% of the currently measured 3×10^6 M$_\odot$ is extended the periastron shifts of the two mechanisms compensate each other. The percentage will be higher for stars on closer orbits of similar or even higher eccentricity.
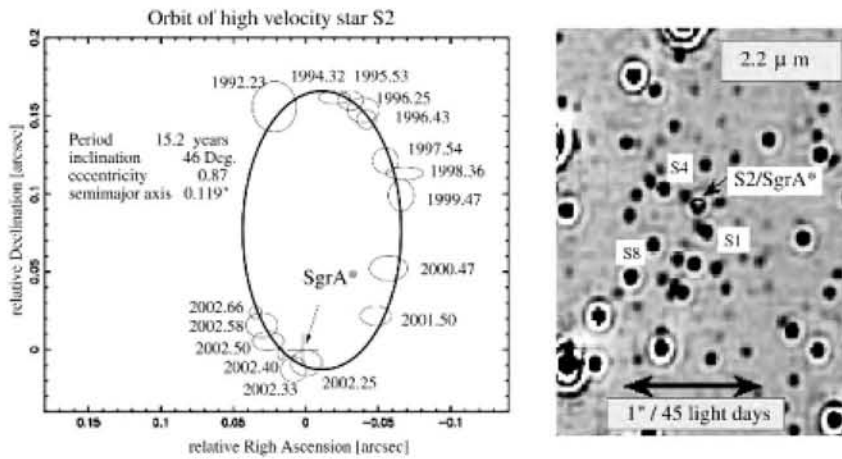
Figure 8.9. Left: measured data points and elements of the high velocity star S2. Right: NACO AO (adaptive optics) image from the central few arcseconds with some of the sources labeled.

Note added in proof

Recently the combination of the 10 years of SHARP measurements at the ESO NTT combined with new VLT UT4 adaptive optics measurements using NACO at the ESO VLT UT4 allowed us (Schödel *et al* 2002) to trace two-thirds of the complete orbit of the star S2, currently closest to the compact radio source and massive black hole candidate Sgr A*. The observations confirm the result by Eckart *et al* (2002) that the star is on a bound, highly elliptical Keplerian orbit around Sgr A*. The orbital period is 15.4 years and the distance during the pericenter passage has been only 17 light hours. The orbital elements require an enclosed point mass of $3.7 \pm 1.5 \times 10^6$ solar masses, which agrees well with the mass at much larger distances. See figure 8.9.

References

Adams D J, Becklin E E, Jameson R F, Longmore A J, Sandquist A A and Valentijn E 1988 *Astrophys. J.* **327** L65

Alexander T and Loeb A 2001 *Astrophys. J.* **551** 223

Alexander T and Sternberg A 1999 *Astrophys. J.* **520** 137

Backer D C 1996 *Proc. of the 169th Symp. of the IAU* p 193

Baganoff F K *et al* 2001a *Astrophys. J.* submitted *preprint* astro-ph0102151

——2001b *Nature* **413** 45

Becklin E E and Neugebauer G 1968 *Astrophys. J.* **151** 145

Blitz L, Binney J, Lo K Y, Bally J and Ho P T P 1993 *Nature* **361** 417

Christou J C 1991 *Exp. Astron.* **2** 27

Close L M, McCarthy D W Jr and Melia F 1995 *Astrophys. J.* **439** 682

Dainty J C (ed) 1975 *Laser Speckle and Related Phenomena* (Berlin: Springer) p 255

Davidge T J, Simons D A, Rigaut F, Doyon R and Crampton D 1997 *Astrophys. J.* **114** 2586

Eckart A and Genzel R 1996 *Nature* **383** 415

——1997 *Mon. Not. R. Astron. Soc.* **284** 576

Eckart A, Genzel R, Hofmann R, Sams B J and Tacconi-Garman L E 1995 *Astrophys. J.* **445** L26

Eckart A, Genzel R, Krabbe A, Hofmann R, van der Werf P P and Drapatz S 1992 *Nature* **335** 526

Eckart A, Genzel R, Ott T and Schödel R 2002 *Mon. Not. R. Astron. Soc.* **331** 917–34

Eckart A, Ott T and Genzel R 1999 *Astron. Astrophys.* **352** 22

Eckart A, Ott T, Genzel R and Rubilar G 2001 *Science with the LBT (Proc. Workshop held at Ringberg Castle, Tegernsee, Germany 24–29 July 2000)* ed T Herbst (Heidelberg: Neumann Druck)

Figer D *et al* 2000 *Astrophys. J.* **533** 49

Forrest W J, Pipher J L and Stein W A 1986 *Astrophys. J.* **301** L49

Fried D L 1966 *J. Opt. Soc. Am.* **56** 1372

Genzel R 1989 *The Center of the Galaxy* ed M Morris (Dordrecht: Kluwer) p 393

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219

Genzel R, Hollenbach D J and Townes C H 1994 *Rep. Prog. Phys.* **57** 417

Genzel R, Pichon C, Eckart A, Gerhard O E and Ott T 2000 *Mon. Not. R. Astron. Soc.* **317** 348

Genzel R, Thatte N, Krabbe A, Kroker H and Tacconi-Garman L E 1996 *Astrophys. J.* **472** 153

Genzel R and Townes C H 1987 *Ann. Rev. Astron. Astrophys.* **25** 377

Genzel R, Watson D M, Crawford M K and Townes C H 1985 *Astrophys. J.* **297** 766

Gerhard O 2000 *Astrophys. J.* **546** L39

Gezari S, Ghez A M, Becklin E E, Larkin J, McLean I and Morris M 2002 *Astrophys. J.* **576** 790

Ghez A M, Klein B L, Morris M and Becklin E E 1998 *Astrophys. J.* **509** 678

Ghez A, Morris M, Becklin E E, Kremenek T and Tanner A 2000 *Nature* **407** 349

Greenhill L *et al* 1995 *Astrophys. J.* **440** 619

Haller J W, Rieke M J, Rieke G H, Tamblyn P, Close L and Melia F 1996 *Astrophys. J.* **456** 194

Herbst T M, Beckwith S V W and Shure M 1993 *Astrophys. J.* **411** L21

Krabbe A *et al* 1995 *Astrophys. J.* **447** L95

Klein B L, Ghez A M, Morris M and Becklin E E 1996 *The Galactic Center (ASP Conference Series 102)* ed R Gredel p 228

Lacy J H, Townes C H, Geballe T R and Hollenbach D J 1980 *Astrophys. J.* **241** 132

Leonard P J T and Merritt D 1989 *Astrophys. J.* **339** 195

Lucy L B 1974 *Astron. J.* **79** 745

Maoz E 1998 *Astrophys. J.* **494** L13

Mariotti J-M 1989 *Diffraction-Limited Imaging with Very Large Telescopes* ed D M Alloin and J-M Mariotti (Dordrecht: Kluwer) p 3

Markoff S, Falcke H, Yuan F and Biermann P L 2001 *Astron. Astrophys.* **379** L13

Melia F and Falcke H 2001 *Annu. Rev. Astron. Astrophys.* **39** 309

Menten K M, Reid M J, Eckart A and Genzel R 1996 *Astrophys. J.* **475** L111

Mezger P G, Duschl W J and Zylka R 1996 *Astron. Astrophys. Rev.* **7** 289

Morris M and Serabyn E 1996 *Annu. Rev. Astron. Astrophys.* **34** 645

Myoshi M *et al* 1995 *Nature* **373** 127

Najarro F, Hillier D J, Kudritzki R P, Krabbe A, Genzel R, Lutz D, Drapatz S and Geballe T R 1994 *Astron. Astrophys.* **285** 573

Najarro F, Krabbe A, Genzel R, Lutz D, Kudritzki R P and Hillier D J 1997 *Astron. Astrophys.* **325** 700

Reid M J, Readhead A C S, Vermeulen R C and Treuhaft R N 1999 *Astrophys. J.* **524** 816

Roddier F 1989 *Diffraction-Limited Imaging with Very Large Telescopes* ed D M Alloin and J-M Mariotti (Dordrecht: Kluwer) p 33

Rubilar G F and Eckart A 2001 *Astron. Astrophys.* **374** 95

Schödel R *et al* 2002 *Nature* **419** 694–6

Sellgren K, McGinn M T, Becklin E E and Hall D N B 1990 *Astrophys. J.* **359** 112

Serabyn E and Lacy J H 1985 *Astrophys. J.* **293** 445

Simon M, Chen W P, Forrest W J, Garnett J D, Longmore A J, Gauer T and Dixon R I 1990 *Astrophys. J.* **360** 95

Simons D A, Hodapp K W and Becklin E E 1990 *Astrophys. J.* **360** 106

Stebbins J and Whitford A E 1947 *Astrophys. J.* **106** 235

Tamblyn P, Rieke G H, Hanson M M, Close L M, McCarthy D W and Rieke M J 1996 *Astrophys. J.* **456** 206

Zhao J-H, Bower G C and Goss W M 2001 *Astrophys. J.* **547** L29

Chapter 9

Stars and singularities: stellar phenomena near a massive black hole

Tal Alexander
The Weizmann Institute of Science, Rehovot, Israel

Besides the Keplerian motion of stars near to the Galactic black hole, described in chapter 8, other detectable effects of black holes on stars and star clusters surrounding it are discussed: stars collide, are tidally disrupted, segregate by mass, and a stellar cusp is formed. The black hole also acts as a gravitational lens, leading to multiple images of individual stars and apparent flares in their luminosity. Lensing can also be used to pinpoint the position of the black hole.

9.1 Introduction

Isolated black holes are simple objects, characterized by three numbers only: mass, angular momentum and charge. The complexity arises from their interaction with their surroundings, which results in a wealth of physical phenomena. This chapter will focus on the interaction of the central $\sim 3 \times 10^6 \, M_\odot$ massive black hole (MBH) in the Galactic Center (Genzel *et al* 2000) with the stars very close to it. We will discuss processes for which there is already some observational evidence, as well as processes that are suggested by theory and may yet be discovered by future observations of the MBH in the Galactic Center or of those in other galaxies.

The MBH environment is unique because in addition to the gravitational singularity, there are three other 'effective singularities' that are associated with the MBH.

(1) A *stellar density singularity*. This is predicted to occur in most scenarios for the evolution of a stellar system around a MBH (e.g. Bahcall and Wolf 1976, 1977, Young 1980). A density distribution that formally diverges at the origin is called a cusp. In practice, infinite density is not reached. Stars

cannot exist closer than the event horizon, and in fact they are destroyed well before that point either by collisions or by the MBH tidal field.

(2) A *velocity singularity*. Close to the MBH the velocity field is Keplerian and so formally diverges as $r^{-1/2}$. The velocity cusp is also limited in practice by the absence of stars arbitrarily close to the MBH and the event horizon.

(3) An *optical singularity*. Any mass bends light and magnifies the flux of background sources. Behind the MBH (or any other sufficiently compact mass) there is a small region (a caustic) where the magnification formally diverges to infinity. This divergence is truncated by the finite size of the source.

The discussion will focus on the consequences of these singularities for stars near the MBH, where the term 'near' is defined here to mean the region where stars can exist (i.e. beyond the tidal radius) but where the potential is completely dominated by the MBH. For the Galactic Center, the event horizon is much smaller than the tidal radius (for a solar-type star) and so general relativistic effects can be neglected to first approximation.

There are several reasons to study stars near the MBH. First, unlike gas, whose dynamics can be influenced by non-gravitational forces such as thermal pressure, radiation pressure and magnetic fields, stars are clean gravity probes. The properties of stars are well known from other environments, and their observed luminosity and spectrum can be translated into mass and maximal age. Both these quantities are very important for understanding the dynamics of the system. In particular, processes that operate on timescales much longer than the maximal stellar age cannot be relevant for the star. Second, stars very near the MBH are connected to the growth of the MBH through tidal disruption, mass loss from stellar winds and from stellar collisions. Third, the region near the MBH can provide a unique laboratory for studying stellar phenomena under extreme conditions: high density, velocity and strong tidal fields.

Presently, infrared spectroscopy is possible for the brighter, well separated stars in the field. Spectroscopy indicates that the stellar population is a mix of old (red) stars and young (blue) stars (see review by Genzel *et al* 1994). The old red giants seen near the MBH in the Galactic Center are in the mass range ~ 1–$8\ M_\odot$ and are older than 1 Gyr. The faintest observable young blue giants may be main sequence B1 or O9 stars with masses of $\sim 20\ M_\odot$ and main-sequence lifetimes <5 Myr. The brightest young stars, the 'He stars', are Wolf–Rayet-like stars with masses of $>20\ M_\odot$ and lifetimes of <10 Myr. The blue stars are too young to have relaxed dynamically, and their orbits (position, velocity) still reflect the initial conditions of their formation (e.g. the young blue emission line giants are observed to rotate in an opposite sense to the Galactic rotation).

All the stars in the inner 0.02 pc around Sgr A* are faint, and have blue featureless spectra, which are typical of young stars. The fact that there are only seemingly young stars very close to the MBH, while there is a mixture of young and old stars farther out, raises a 'Nature versus Nurture' question.

Is this an essentially random variation in the stellar population, which can be explained in terms of normal star formation processes ('Nature'), or is this a result some systematic effects of the unique extreme environment very near the MBH ('Nurture')? It is interesting to note that a cluster of blue stars exists also around the $\sim$3 $\times$ 10^7 $M_\odot$ MBH in the galaxy M31 (Lauer *et al* 1998).

If these stars are indeed the products of their environment, then there are two options to consider. First, this could be the result of an unusual mechanism of star formation (Morris 1993 and chapter 4), in which case the stars are indeed young and dynamically unrelaxed, and so do not convey direct information on the dynamical processes near the MBH. Second, this could be the result of unusual stellar evolution, so that the stars only appear young, but are in fact old and dynamically relaxed. This chapter will focus on the second possibility (section 9.3). However, before we discuss possible mechanisms for modifying stellar evolution, it is useful to review some results from stellar dynamics theory.

9.2   Stellar dynamics near a black hole

The stellar dynamical term 'stellar collision' is not limited to the case of actual physical contact between stars, but refers to any gravitational interaction where the stars exchange momentum or energy. The dynamical processes in a gravitating stellar system can be summarized by classifying stellar collisions according to their distance scale. The reader is referred to Binney and Tremaine (1987) for a detailed treatment of this subject.

On the largest scale, the motion of a star is determined by the sum of interactions with all the other stars, that is, by the smooth gravitational potential of the system. Two-body interactions occur on a shorter length scale, when two stars approach each other to the point where their mutual interaction dominates over that of the smoothed potential. Two-body interactions randomize the stellar motions and lead to the relaxation of the system. In the course of relaxation, the stars, whose mass range spans two to three orders of magnitude, are driven towards equipartition. However, equipartition cannot be achieved in the presence of a central concentration of mass (in particular a central MBH). When two stars, which are initially on the same orbit (and therefore have the same velocity) interact, the massive one will slow down and the lighter one will speed up. Since the radius of the orbit depends only on the star's *specific* energy, and not its total energy, the massive star will sink to the center, while the lighter star will drift outwards. Over time, this process leads to 'mass segregation'—the more massive stars are concentrated near the MBH and the lighter stars are pushed out of the inner region.

Occasionally, two-body interactions will eject a star out of the system altogether, thereby taking away positive energy from the system. The system will then become more bound and compact, the collision rate will increase, more stars will be ejected, and the result will be a runaway process. This process is

called the 'gravothermal catastrophe', or 'core collapse', and is linked to the fact that self-gravitating systems have a negative heat capacity—they become hotter when energy is taken out. Core collapse, if unchecked, will lead to the formation of an extremely dense stellar core surrounded by a diffuse extended halo.

Once the density becomes high enough, very short-range inelastic collisions are no longer extremely rare, and the fact that the stars are not point masses but have internal degrees of freedom starts to play a role. In such collisions energy is extracted from the orbit and invested in the work required to raise stellar tides, or strip stellar mass. The tidal energy is eventually dissipated in the star and radiated away. If the collision is slow, as it is in the core of a globular cluster where there is no MBH (but see Gebhardt *et al* 2002 for evidence for a black hole in a globular cluster), then the typical initial orbit is just barely unbound. In this case, the tidal interaction may extract enough orbital energy for 'tidal capture', and lead to the formation of a tightly bound, or 'hard' binary (tight, because tidal forces become effective only when the two stars are very close to each other). Hard binaries are a heat source for the cluster and play a crucial role in arresting core collapse. When a third star collides with a hard binary, it will tend to gain energy from the binary, thereby injecting positive energy into the cluster, while the binary becomes harder still.

When the stars orbit a central MBH, the collisions are fast (the Keplerian velocity near the MBH exceeds the escape velocity from the star) and the initial orbits are very unbound (hyperbolic). Even very close fly-bys cannot take enough energy from the orbit to bind the two stars, and so they continue on their way separately after having extracted energy and angular momentum from the orbit. The stars can radiate away the excess heat on a timescale shorter than the mean time between collisions, but it is harder to get rid of the excess angular momentum. Magnetic braking (the torque applied to a star when the stellar wind resists being swept by the rotating stellar magnetic field) typically operates on timescales similar to the stellar lifetime. It is therefore likely that high rotation is the longest-lasting dynamical after effect of a close hyperbolic encounter, and that stars in a high density cusp are spun up stochastically by repeated collisions (section 9.3.2). Finally, at zero range, almost head-on stellar collisions can lead to the stripping of stellar envelopes (section 9.3.1), the destruction of stars, or to mergers that result in the creation of 'exotic stars'. These are stars that cannot be formed in the course of normal stellar evolution, such as a Thorne–Zytkow object, which is an accreting neutron star embedded in a giant envelope (Thorne and Zytkow 1975).

9.2.1 Physical scales

There are several important timescales and length scales that govern the dynamics of the stellar system and MBH. They are listed here with estimates of their value in the Galactic Center. A solar-type star and $M_\bullet = 3 \times 10^6 \ M_\odot$ (Genzel *et al* 2000) are assumed throughout. Physical lengths are also expressed as angular sizes assuming that the distance to the Galactic Center is $R_0 = 8$ kpc (Reid 1993).

9.2.1.1   Timescales

The dynamical time, or orbital time, t_d, is the time it takes a star to cross the system

$$t_d \sim \frac{r}{v} \sim 2\pi \sqrt{\frac{r^3}{GM_{tot}}} \sim 2 \times 10^5 \text{ yr (at 3 pc)} \sim 300 \text{ yr (at 0.03 pc)} \quad (9.1)$$

where r is the typical size of the system and M_{tot} is the total mass enclosed in radius r.

The two-body relaxation time, t_r, is related to the 1D velocity dispersion σ, the mean stellar mass $\langle M_\star \rangle$ and the stellar number density $n_\star$ by

$$t_r \sim \frac{0.34\sigma^3}{G^2 \langle M_\star \rangle^2 n_\star \ln \Lambda} \sim 10^9 \text{ yr} \quad (9.2)$$

where $\log \Lambda$ is the Coulomb logarithm, the logarithm of the ratio between the largest and smallest impact parameters possible in the system for elastic collisions. Because the relaxation timescale in the Galactic Center is shorter than the age of the Galaxy (~ 10 Gyr), the old stars are expected to be well relaxed by now.

The mass segregation timescale is of the same order as the relaxation timescale,

$$t_{seg} \sim t_r. \quad (9.3)$$

The rate (per star) of grazing collisions between two stars of mass and radius $M_\star^a$, $R_\star^a$ and $M_\star^b$, $R_\star^b$, each, is

$$t_c^{-1} = 4\sqrt{\pi} n_\star \sigma (R_\star^a + R_\star^b)^2 \left[1 + \frac{G(M_\star^a + M_\star^b)}{2\sigma^2(R_\star^a + R_\star^b)} \right] \sim 10^{-9} \text{ yr}^{-1} \text{ (at 0.03 pc)} \quad (9.4)$$

where it is assumed that the stars follow a mass-independent Maxwell–Boltzmann velocity distribution with velocity dispersion σ (this is a good approximation near the MBH, see section 9.2.2). There are two contributions to the total rate, one due to the geometric cross section (first term in the square brackets) and one due to 'gravitational focusing' (second term in the square brackets). Gravitational focusing expresses the fact that the two stars do not move on straight lines, but are attracted to each other. This effect is important when the typical stellar velocities are much smaller than the escape velocity, v_e, from the stars, $\sigma^2 < GM_\star/2R_\star = v_e^2/4$.

9.2.1.2   Length scales

The size of the event horizon of a non-rotating black hole, the Schwarzschild radius, is

$$r_S = \frac{2GM_\bullet}{c^2} = 9 \times 10^{11} \text{ cm} \sim 3 \times 10^{-7} \text{ pc} \sim 8 \text{ } \mu\text{arcsec.} \quad (9.5)$$

The tidal radius, r_t, is the minimal distance from the MBH where the stellar self-gravity can still resist the tidal forces of the MBH. If the star's orbit takes it inside the tidal radius, it will be disrupted, and roughly half of its mass will fall into the MBH, while the other half will be ejected (e.g. Ayal *et al* 2000). The exact value of the tidal radius depends on the stellar structure and the nature of the orbit, and up to a factor of order unity is given by

$$r_t \sim R_\star \left(\frac{M_\bullet}{M_\star} \right)^{1/3} = 10^{13} \, \text{cm} \sim 3 \times 10^{-6} \, \text{pc} \sim 80 \, \mu\text{arcsec.} \qquad (9.6)$$

Tidal disruption is relevant as long as the tidal radius lies outside the event horizon. Since $r_t \propto M_\bullet^{1/3}$, while $r_S \propto M_\bullet$, there exists a maximal MBH mass for tidal disruption, which for a solar-type star is $\sim 10^8 \, M_\odot$.

The radius of influence, r_h, is the region where the MBH potential dominates the dynamics. If the MBH is embedded in an isothermal stellar system (i.e. σ is constant), then the radius of influence can be defined as

$$r_h = \frac{GM_\bullet}{\sigma^2} \sim 10^{19} \, \text{cm} \sim 3 \, \text{pc} \sim 80 \, \text{arcsec.} \qquad (9.7)$$

In practice, the distribution is not isothermal and σ is not constant, and so r_h is evaluated loosely by choosing a representative value of σ far enough from the MBH. The stellar mass enclosed within r_h is of the same order as the mass of the MBH.

9.2.2 A relaxed stellar system around a MBH

The relaxed, quasi steady-state density distribution of a single-mass stellar population around a MBH is (Bahcall and Wolf 1976, see also Binney and Tremaine 1987 for a simple derivation)

$$n_\star \propto r^{-7/4}. \qquad (9.8)$$

When the stellar population consists of a spectrum of masses, $M_1 < M_\star < M_2$, the stellar distribution function (DF) very near the MBH has the form (Bahcall and Wolf 1977)

$$f_M(\epsilon) \propto \epsilon^{p_M} \qquad n_\star \propto r^{-3/2 - p_M} \qquad p_M \equiv \frac{M}{4M_2} \qquad (9.9)$$

where $-\epsilon$ is the total specific energy of the star and $f_M \equiv 0$ for $\epsilon < 0$. The velocity dispersion of this DF (see equation (9.12)) is almost independent of the stellar mass,

$$\sigma_M^2 = \left(\frac{1}{5/2 + p_M} \right) \frac{GM_\bullet}{r} \qquad (9.10)$$

which implies that σ_M^2 changes by less than 10% over the entire mass range, in marked contrast to the $\sigma_M^2 \propto M_\star^{-1}$ dependence of equipartition. This result
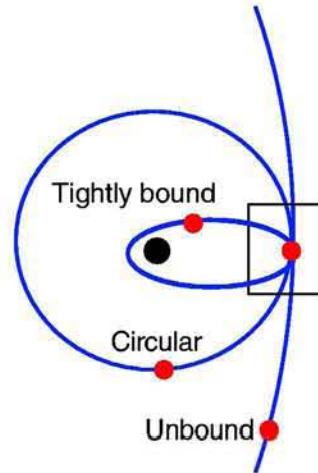
Figure 9.1. A schematic representation of the different types of stellar orbit that can be observed in a small region near the MBH.

justifies the approximation that the velocity dispersion in a relaxed stellar system around a MBH is mass independent.

The Bahcall–Wolf solution applies to point particles. This assumption no longer holds very near the MBH, where the collision rate is high because of the very high stellar density. Stars on tight orbits around the MBH cannot survive for long, and so eventually most of the population there will consist of stars that are on very wide, marginally bound (parabolic) orbits, which spend only a small fraction of their time in the collisionally dominated region. These marginally bound stars have a flatter spatial distribution, of the form (e.g. Binney and Tremaine 1987, p 551)

$$n_\star \propto r^{-1/2}. \tag{9.11}$$

The stars in any volume element near the MBH have a distribution of orbits (figure 9.1): some are more bound than circular (i.e. their velocity is smaller than the circular velocity v_c), some are less bound than circular, some are unbound to the MBH (but bound by the total mass of the MBH and stars). The distribution of orbits is directly tied to the spatial distribution through the Jeans equation,

$$\frac{GM_\bullet}{r\sigma^2} = \frac{v_c^2}{\sigma^2} = -\frac{d\ln n_\star}{d\ln r} - \frac{d\ln\sigma^2}{d\ln r}. \tag{9.12}$$

The Jeans equation is essentially a re-statement of the continuity equation of the stellar orbits in phase space in terms of averaged quantities, the mean stellar density and velocity dispersion. Here it is given for the simplest case of a steady-state, isotropic, non-rotating system. The steady-state assumption is justified
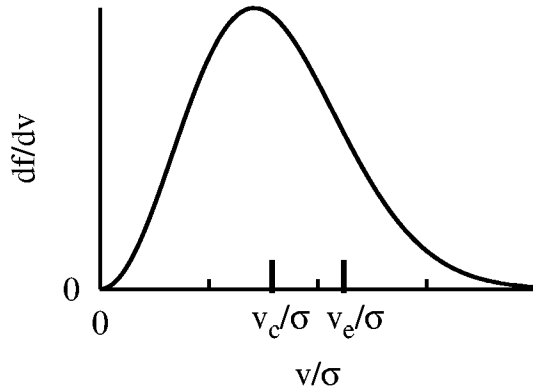
Figure 9.2. The fraction of orbits in the Maxwell–Boltzmann distribution as a function of the normalized velocity v/σ. The circular velocity v_c and the escape velocity $v_e = \sqrt{2}v_c$ are marked for a density distribution with $\alpha = 1.5$. The region $v > v_e$ lies under the exponential tail of the DF, and so the fraction of stars with unbound orbits is a strongly decreasing function of α.

because the dynamical timescale is much shorter than the relaxation timescale. The assumptions of approximate isotropy and non-rotation are observationally justified.

Very near the MBH the velocity dispersion is Keplerian, $\sigma^2 \propto r^{-1}$, and so for any power-law cusp $n_\star \propto r^{-\alpha}$, the Jeans equation implies that

$$\frac{v_c^2}{\sigma^2} = \alpha + 1. \tag{9.13}$$

The steeper the cusp (larger α) is the larger the ratio between v_c and σ, and so the fraction of loosely bound stars or unbound stars is smaller (figure 9.2). Because unbound stars have wide orbits and spend most of their time far away from the MBH, the stellar population in a shallow cusp is well mixed and representative of the average population over a large volume. In contrast, the stellar population in a steep cusp is localized and can therefore develop and maintain properties that differ from those of the general population.

9.3 The stellar collider in the Galactic Center

The potential for probing a new regime of stellar dynamics near the MBH in the Galactic Center is best illustrated by comparing the collisional timescale there with that in the cores of the densest globular clusters, which for a long time served as laboratories for the study of collisional processes. In a dense globular

cluster, $n_\star \sim 10^6 \ M_\odot \ \text{pc}^{-3}$ and $\sigma \sim 10 \ \text{km s}^{-1}$, whereas in the Galactic Center, the density may be as high as $n_\star \sim 10^8 \ M_\odot \ \text{pc}^{-3}$ (section 9.3.1) and $\sigma \sim 1000 \ \text{km s}^{-1}$. The timescale for collisions between solar-type stars in a globular cluster can be estimated from equation (9.4) to be almost 10^{10} yr, roughly the age of the Galaxy and of a solar-type star, whereas it is only $\sim 5 \times 10^8$ yr in the inner 0.02 pc of the Galactic Center. These estimates imply that while physical collisions are only marginally relevant in the cores of the densest globular clusters, they completely dominate the dynamics in the innermost part of the MBH cusp[1].

9.3.1   The case for a dense stellar cusp in the Galactic Center

Theoretical considerations lead us to expect a relaxed stellar cusp around the MBH in the Galactic Center. Does such a cusp indeed exist there? The answer depends critically on the problem of identifying which of the observed stars are dynamically relaxed, since only those faithfully trace the underlying old stellar population. The analysis presented here shows that it is possible to interpret the available observations *self-consistently* in the framework of a high density cusp. However, the reader should keep in mind that the issue is an empirical one and, as such, may be subject to revisions when more and better data are obtained about the stars near the MBH.

Direct evidence for the existence of a cusp comes from the analysis of star maps, which show a concentration of stars toward the center. Assuming a 3D density distribution of the form $n_\star \propto r^{-\alpha}$, the corresponding projected 2D surface density can be compared to the observed distribution to find the most likely value of α. Figure 9.3 shows the likelihood curves for α based on three independent star maps, after all the stars that were spectroscopically identified as young were taken out of the sample (the faint blue stars nearest to Sgr A* are included only in the Keck data set, but not in the other two). The most likely value for the density power-law index α lies in the range ~ 1.5–1.75. A flat core ($\alpha \sim 0$), such as exists in globular clusters, is decisively rejected. Similarly, a likelihood test for the maximal size of a flat inner core indicates that such a core, if it exists, is smaller than ~ 0.1 pc (2.5″). It can be shown that extinction by interstellar dust is unlikely to bias these results by a significant amount.

Additional evidence for the existence of a very high density cusp comes from the observed gradual depletion of the luminous giants toward the MBH in the inner 0.1 pc (figure 9.4). Luminous red giants have very large extended envelopes, and therefore a large cross section for collisions with other stars. When the impact parameter is a small fraction of the giant's radius, the envelope may be stripped, leaving behind an almost bare burning core. This will make the star effectively invisible in the infrared (IR) because the IR spectral range lies in the Rayleigh–Jeans part of the stellar blackbody spectrum, and so the IR luminosity scales as $L_{\text{IR}} \propto R_\star^2 T_{\text{eff}}$ while the total luminosity scales as $L_\star \propto R_\star^2 T_{\text{eff}}^4$. Suppose that

---

[1] The probability for *avoiding* a collision over a time t is $\exp(-t/t_{\text{c}})$.
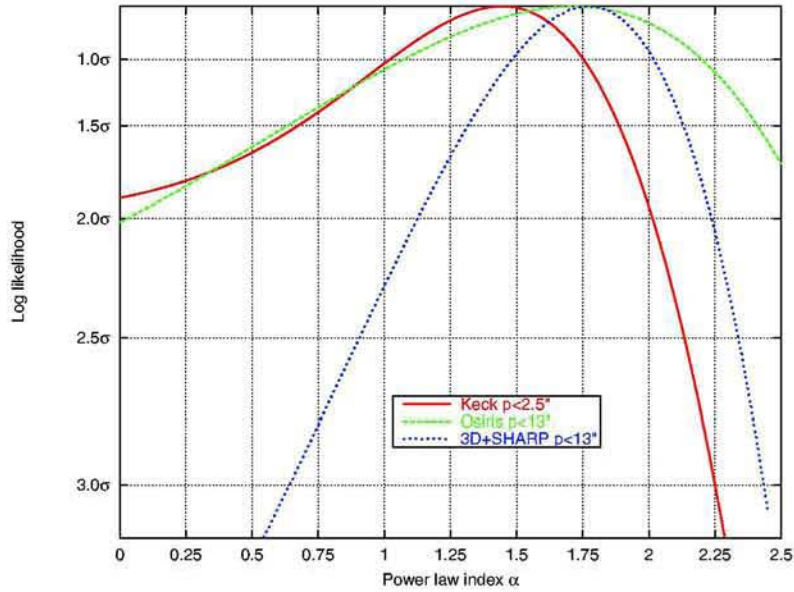
Figure 9.3. A maximum likelihood analysis of the surface density distribution of stars near Sgr A* for a 3D stellar density distribution $n_\star \propto r^{-\alpha}$ (Alexander 1999). Three different data sets (Blum *et al* 1996, Genzel *et al* 1996, Eckart and Genzel 1997, Ghez *et al* 1998) indicate that the most likely value for α lies in the range $\sim 3/2$ to $\sim 7/4$, which is the theoretically predicted range for a relaxed stellar system around a MBH (Bahcall and Wolf 1977). Order of magnitude estimates (section 9.2.1) suggest that the stellar system around the MBH in the Galactic Center has undergone two-body relaxation. (Reprinted with permission from *The Astrophysical Journal*.)

the collision disperses the envelope of a $\sim 100\ R_\odot$ red supergiant and leaves a $\sim 1\ R_\odot$ burning core. In order to maintain the total stellar luminosity, the effective temperature will have to rise by a factor of 10, which will result in a decrease of the IR luminosity by a factor of 1000 (7.5 magnitudes).

Figure 9.4 compares a theoretical prediction for the collisional depletion of luminous giants with the data. The match with the observed trend is remarkably good, given the fact that no attempt was made to fit the data. The calculation is based on detailed modeling of expected numbers, sizes, luminosities and lifetimes of giants in the population, on cross sections for envelope disruption that were calibrated by hydrodynamical simulations, and on a stellar density cusp that is normalized by dynamical estimates of the enclosed mass.

It should be noted that the total mass loss rate from these collisions is smaller than that supplied by the strong stellar winds of the blue supergiants in the inner few arcseconds, and so stellar collisions are not a dominant source of mass supply to the MBH at this time.
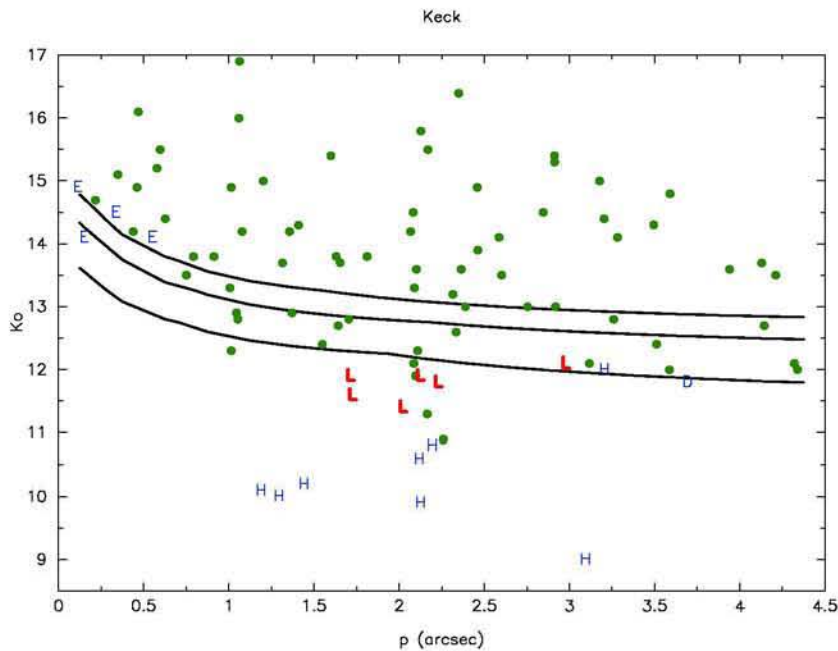
Figure 9.4. Evidence for collisional destruction of bright giant envelopes in a high-density stellar cusp around the MBH in the Galactic Center (Alexander 1999). The apparent stellar K-band magnitude is plotted against the projected angular distance from the black hole, p (Keck data from Ghez *et al* 1998). The ages of the stars marked by circles are unknown, but it is likely that most of them are old, and therefore dynamically relaxed. Stars marked by 'L' are spectroscopically identified as old stars. Stars marked by 'H' are spectroscopically identified as young stars and are not dynamically relaxed. Such stars are not expected to be affected by collisions because of their short lifetimes. The stars marked by 'E' have featureless blue spectra and are either young stars or old stars that were affected by the extreme conditions very near the black hole. The three contour lines represent detailed model predictions for the decrease in surface density of bright stars due to collisional destruction in a high density $n_\star \propto r^{-3/2}$ stellar cusp. The stellar density reaches a value of $\sim 4 \times 10^8 \, M_\odot \, pc^{-3}$ at $r = 0.25''$ (0.01 pc), which is nine orders of magnitude higher than in the Solar Neighborhood, and almost three orders of magnitude higher than in the densest globular cluster core. The model predicts, on average, 1.5 (top contour), 1.0 (central contour), and 0.5 (bottom contour) dynamically relaxed stars per 0.25 arcsecond bin that are brighter than the contour level. This is consistent with the observed trend in the surface density distribution. (Reprinted with permission from *The Astrophysical Journal*.)

The self-consistent picture that emerges from this analysis is that the stars near the MBH in the Galactic Center, which are expected to be dynamically

relaxed, are indeed concentrated in a stellar cusp of the form predicted by theory for a relaxed system. The very high stellar density in the inner few 0.01 pc leads to frequent collisions that destroy the envelopes of giant stars, thereby explaining the gradual depletion in the number of luminous giants toward the center. The central cluster of faint blue stars in the inner $0.5''$ coincides with the collisionally dominated region. It is therefore relevant to consider dynamical explanations for their nature and appearance as an alternative to assuming that they are newly formed, unrelaxed stars. The concentration of such a distinct population in a small volume is consistent with the tightly bound orbits that are typical of a steep cusp (section 9.2.2).

9.3.2 Tidal spin-up

It is inevitable that in a system where the stellar density is high enough for collisional destruction of giants, smaller stars that escape destruction will still suffer very close encounters. As described earlier (section 9.2), usually such collisions cannot bind the two stars, and the longest lasting after effect, apart from possible mass loss, is fast rotation. Fast rotation and mass loss have the potential to affect stellar evolution and modify the appearance of the stars (see the discussion in Alexander and Kumar 2001, Alexander and Livio 2001). Although detailed predictions of the observational consequences are still not available, it is of interest to estimate the magnitude of the spin-up effect.

When the tidal deformations in the star are small, the change in the angular velocity of a star of mass $M_\star$ and radius $R_\star$ due to an encounter with a mass M can be described by a linear multipole expansion in the periapse distance $r_{\rm p}$ (distance of closest approach) by (Press and Teukolsky 1977)

$$\Delta\widetilde{\Omega} = \frac{\widetilde{M}^2}{\widetilde{I}\widetilde{v}_{\rm p}} \sum_{l=2}^{\infty} \frac{T_l(\eta, e)}{\widetilde{r}_{\rm p}^{2l+1}} \tag{9.14}$$

where the tilde symbol denotes quantities measured in units where $G = M_\star = R_\star = 1$, and rigid body rotation is assumed. $\widetilde{v}_{\rm p}$ is the relative velocity at periapse, $\widetilde{I}$ is the star's moment of inertia, and T_l the tidal coupling coefficient of the lth moment. In these units, $\widetilde{\Omega} = 1$ is the centrifugal breakup angular velocity, where the star sheds mass from its equator. The tidal coupling coefficients depend on the star's structure, on the orbital parameters through the quantity $\eta = \widetilde{r}_{\rm p}^{3/2}/\sqrt{1 + \widetilde{M}}$, and on the orbital eccentricity e. The tidal coefficient T_l can be calculated numerically for any given stellar model and orbit.

The formal divergence of $\Delta\widetilde{\Omega}$ as $\widetilde{r}_{\rm p}$ decreases indicates that most of the contribution comes from close collisions, where the linear analysis breaks down. The nonlinear processes, which truncate the divergence, have to be investigated by hydrodynamical simulations (see section 9.3.3). These reveal that as $\widetilde{r}_{\rm p}$ decreases towards 1, $\Delta\widetilde{\Omega}$ first increases faster than predicted by the linear analysis, but then
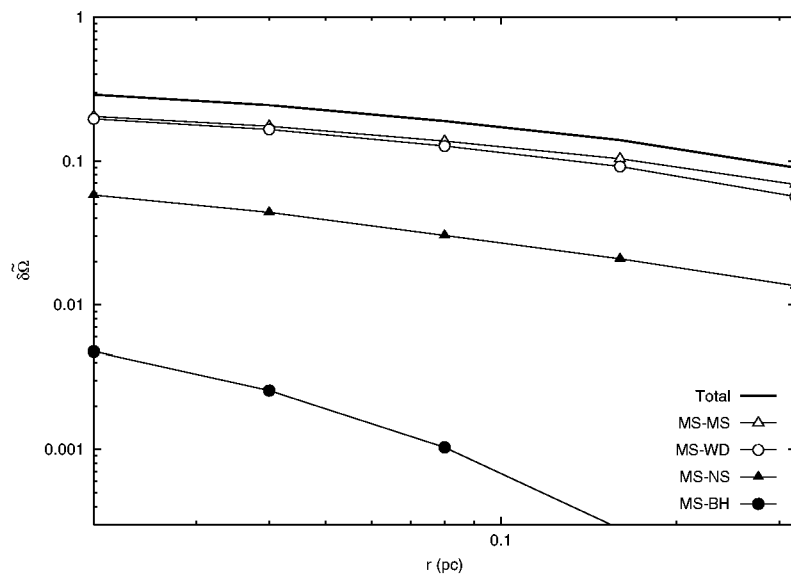
Figure 9.5. The average spin-up of a solar-type star by star–star tidal interactions over 10 Gyr as a function of distance from the black hole in the Galactic Center (Alexander and Kumar 2001). An $\alpha = 1.5$ density cusp is assumed. The rotation grows over time in a random walk fashion by repeated close passages. $\widetilde{\delta\Omega} = 1$ corresponds to rotation at the centrifugal break-up velocity. In addition to the total spin-up, the separate contributions from collisions with main sequence stars (MS), white dwarfs (WD), neutron stars (NS) and stellar black holes (BH) are shown. (Reprinted with permission from *The Astrophysical Journal*.)

it reaches a maximal value at the onset of mass loss, since the ejecta carry away the extra angular momentum.

Over its lifetime, a star will undergo many tidal encounters, randomly oriented relative to its spin axis, and will be spun up in a random walk manner. The cumulative effect can be large. Figure 9.5 shows the predicted average spin-up of solar-type stars over 10 Gyr in the Galactic Center as a function of distance from the MBH (Alexander and Kumar 2001). The calculation assumes an $\alpha = 1.5$ density cusp, a model for the distribution of stellar masses in the population, and inefficient magnetic braking. On average, solar-type stars in a large volume around the black hole are spun up to 10%–30% of the break-up angular velocity, or 20 to 60 times faster than is typical in the field. The effect falls off only slowly with distance because the higher efficiency of tidal interactions in slower collisions far from the black hole offsets the lower collision rate there.

9.3.3 Tidal scattering

Tidal scattering is another mechanism that can affect the internal structure of a significant fraction of the stars around the MBH. Unlike the tidal spin-up process discussed in section 9.3.2, tidal scattering does not require a very high stellar density, since it is driven by the global response of the system to the existence of a mass sink, the MBH, in its center.

Some of the mass that feeds the growth of a MBH in a galactic center is supplied by tidal disruption of stars that are scattered into low angular momentum orbits ('loss-cone' orbits). When the MBH mass is small enough that the tidal radius is larger than the event horizon, $r_t > r_s$, the star is tidally disrupted before crossing the event horizon. The accretion of stellar debris from such events may give rise to observable 'tidal flares' (Frank and Rees 1976). Significant theoretical efforts have gone into estimating the rates, timescales, luminosities and spectra of the flares (e.g. Ulmer *et al* 1998, Magorrian and Tremaine 1999, Ayal *et al* 2000), in the hope that they can be used to detect MBHs in the centers of galaxies. There is today only marginal evidence for the detection of such flares (e.g. Renzini *et al* 1995, Komossa and Bade 1999, Komossa and Greiner 1999).

The effect of the MBH's tidal field is not limited to tidal disruption. For every star that is actually disrupted, there are stars with $r_p \gtrsim r_t$ that narrowly escape tidal disruption by the central BH after being subjected to extreme tidal distortion, spin-up, mixing and mass loss, which may affect their evolution and appearance (Alexander and Livio 2001). Figure 9.6 shows a Smoothed Particle Hydrodynamics[2] (SPH) simulation of a star passing by a MBH just outside the tidal disruption radius. To leading order, the effects of tidal scattering are a function of the penetration parameter $\beta = r_t/r_p$ only, and are independent of the MBH mass,

$$\Delta\widetilde{\Omega} \simeq \frac{T_2(\beta^{-3/2})}{\sqrt{2\widetilde{I}}}\beta^{9/2} \tag{9.15}$$

which follows from equation (9.14) for a parabolic orbit and for $M_\bullet \gg M_\star$. As will be argued later, a large fraction of these 'tidally scattered' stars survive eventual orbital decay and disruption, and so remain in the system as relics of the epoch of tidal processes even after the MBH becomes too massive for tidal disruption.

Dynamical analyses of the scattering of stars into the loss-cone orbits (Lightman and Shapiro 1977, Magorrian and Tremaine 1999) show that tidally disrupted stars in galactic nuclei are typically on slightly unbound orbits relative to

---

[2] SPH is a numerical algorithm for simulating the hydrodynamics of 3D self-gravitating fluids, which is commonly used in the study of stellar collisions (Monaghan 1992). The star is represented by discrete mass elements, each distributed smoothly over a small sphere so that its density peaks in the center and falls to zero at the edge. The total density at a point is the sum of densities in all overlapping spheres that include the point. The resulting density field is continuous and differentiable, and so its thermodynamic properties can be evaluated everywhere once an equation of state is specified. Every time step, the positions of the mass elements are updated according to the gravitational force and the pressure gradient, and the sphere sizes are readjusted to reflect the changes in the local density.
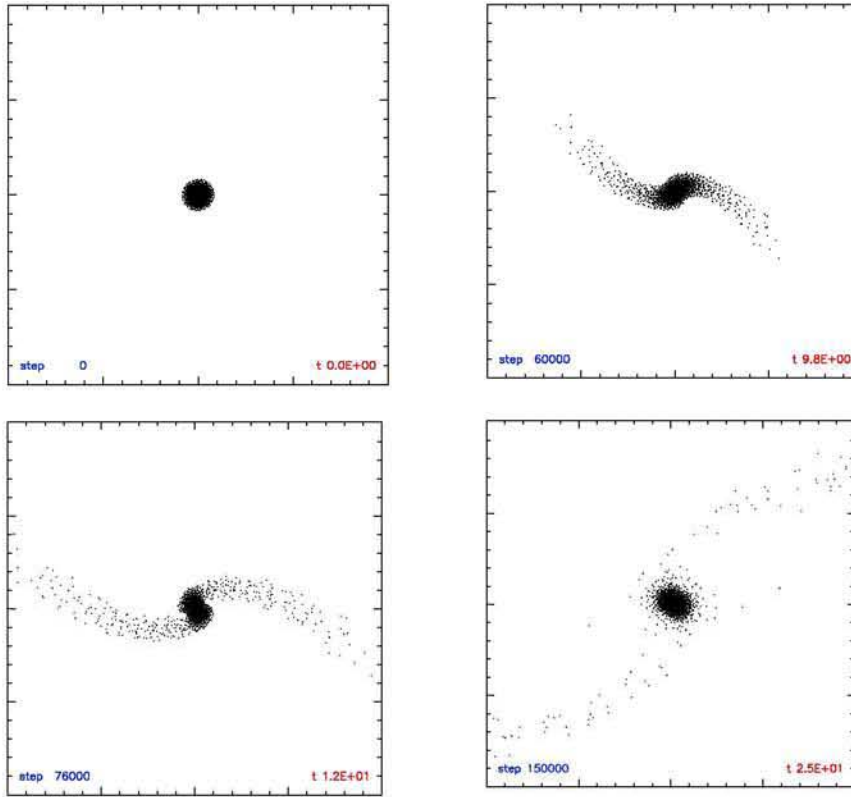
Figure 9.6. Snapshots from a Smoothed Particle Hydrodynamics (SPH) simulation of a star undergoing an extreme non-disruptive tidal interaction ('tidal scattering') as it passes near a massive black hole. Time is measured in units of the star's dynamical timescale. The star passes near the black hole (located outside the frame) on a parabolic orbit with a peri-distance 1.5 times larger than the tidal disruption distance. Shortly after periapse passage ($t = 12$) the star appears to be on the verge of breaking in two. However, by the end of the simulation, the two fragments coalesce, leaving a distorted, mixed and rapidly rotating bound object.

the MBH and that they are predominantly scattered into the loss-cone from orbits at the radius of influence of the BH, r_h. The scattering operates on a timescale that is shorter than the dynamical timescale, and so the stars are scattered in and out of the loss-cone several times during one orbital period. Because of gravitational focusing, the cross section for scattering into a hyperbolic orbit with periapse $\leq r_p$ scales as r_p, and not as r_p^2 (Hills 1975, Frank 1978), and so the number of stars with $r_t \leq r_p \leq 2r_t$ equals the number of stars that were disrupted by the MBH.

Tidal disruption is an important source of mass for a low-mass MBH that

accretes from a low-density galactic nuclear core, where mass loss from stellar collisions is small (e.g. Murphy *et al* 1991). For the MBH in the Galactic Center, the total mass in disrupted stars can be $0.25M_\bullet$ or even higher (Freitag and Benz 2002). Since the enclosed stellar mass within r_h is also $\sim M_\bullet$, the tidally scattered stars comprise a significantly high fraction of the stellar population within the radius of influence of the MBH.

After the first periapse passage, the tidally scattered star will be on a very eccentric orbit with a maximal radius (apoapse) of $\lesssim 2r_h$. Since the two-body interactions that scattered it into the eccentric orbit operate on a timescale that is shorter than the orbital period, there is a significant chance that the star will be scattered again off the orbit and miss the MBH. The chance of this happening is further increased by the Brownian motion of the MBH relative to the dynamical center of the stellar system. The amplitude of the Brownian motion is much larger than the tidal radius, and it proceeds on the dynamical timescale of the core (Bahcall and Wolf 1976), which is comparable to the orbital period of the tidally disturbed stars. The orbits of the tidally scattered stars take them outside of r_h, where they are no longer affected by the relative shift between the BH and the stellar mass. Therefore, on re-entry into the volume of influence, their orbit will not bring them to the same periapse distance from the MBH (figure 9.7). Both the random motion of the MBH and the scattering off the loss-cone by two-body interactions are expected to increase the survival fraction to a significantly high value. More detailed calculations, which integrate over the orbital distribution, are required to confirm these qualitative arguments.

Rough estimates (Alexander and Livio 2001) indicate that the Galactic Center may harbor 10^{4-5} tidally scattered stars. These stars are expected to be on highly eccentric orbits, and so there may be observable correlations between high orbital eccentricity and the stellar properties.

9.4 The gravitational telescope in the Galactic Center

The MBH in the Galactic Center is a telescope with a lens of effective diameter $\sim 4 \times 10^{17}$ cm (for a source at infinity) and a focal length of $\sim 2.5 \times 10^{22}$ cm. Unfortunately, Nature did not design it as an ideal telescope. A point mass lens does not produce faithful images of the lensed sources, the optical axis is heavily obscured by interstellar dust, and the telescope points in a fixed direction, which is not of our choosing. In fact, various estimates suggest that there are not enough luminous sources in that direction for gravitational lensing to be important for present day observations, although future, deep observations may pick up lensing events (Wardle and Yusef-Zadeh 1992, Alexander and Sternberg 1999, Alexander and Loeb 2001). Nevertheless, it is worthwhile to consider the possible roles of gravitational lensing in the observations and the study of the Galactic Center. This is important not only in anticipation of future observations, but also because the estimates of the lensing probability are quite uncertain (they involve models of
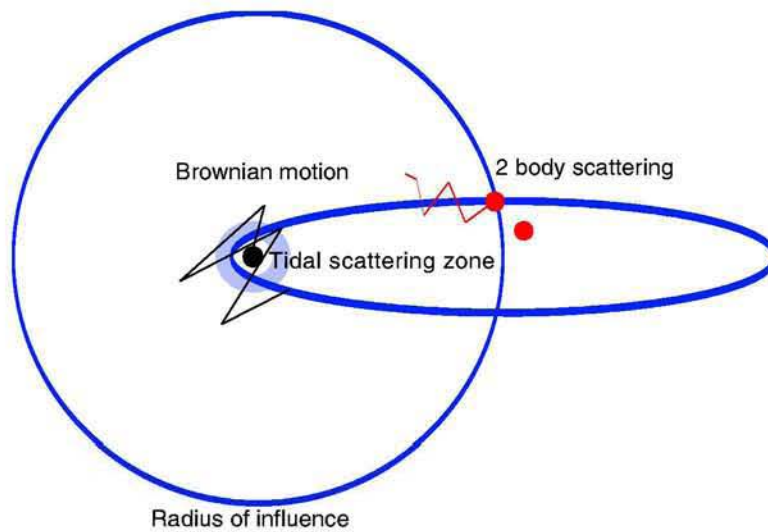
Figure 9.7. A schematic representation of the tidal scattering process. A star initially orbiting the MBH at the radius of influence is scattered by a two-body encounter into an extremely eccentric orbit that brings it to the tidal scattering zone just outside the tidal disruption radius. The star suffers an extreme, non-disruptive tidal interaction with the MBH, and continues on its way out of the radius of influence, where it is scattered by frequent two-body encounters. In the meanwhile, the Brownian motion of the MBH due to its interactions with the stellar system causes it to move away from its original position. Both these random processes significantly increase the chances of the tidally disturbed star surviving total disruption during subsequent orbits.

the unobserved far side of the Galaxy), and because there are hints that lensing may not be quite as rare as predicted (section 9.4.2).

Gravitational lensing may be used to probe the dark mass (is it really a MBH?) and the stars around it, and to locate the MBH on the IR grid, where the stars are observed. However, gravitational lensing can also complicate the interpretation of the observations since it affects many of the observed properties of the sources: flux, variability, apparent motion and surface density. IR flares due to lensing can be confused with those due to fluctuations in the accretion flow, and lensed images of background sources far behind the MBH can be confused with stars that are truly near the MBH. This section will focus on aspects of gravitational lensing that are or may be relevant for the Galactic Center. The reader is referred to Schneider *et al* (1992) for a comprehensive treatment of the subject.

9.4.1 Gravitational lensing by a point mass

To a first and good approximation the lensing properties of the mass distribution in the Galactic Center can be described as those of a point mass, the MBH. Figure 9.8 shows the light ray diagram of lensing by a point mass in the small-angle limit. The bending angle is given by

$$\alpha = \frac{4GM_\bullet}{c^2 b} \tag{9.16}$$

where b is the impact parameter of the light ray with respect to the lens. Note that unlike a spherical glass lens, where the bending angle is zero when the ray goes through the lens center and increases with the impact parameter, the bending angle of a gravitational lens diverges toward the center and decreases with the impact parameter. It is therefore not surprising that a gravitational lens does not produce a faithful image of the lensed source, but rather breaks, warps and/or flips the image. A point lens creates two images of the source, one on either side of the lens. There are always two images in focus at the observer, regardless of the distance of the source behind the lens. The two images, the lens and the (unobserved) source all lie on one line. The typical angular cross section of the lens is given by the Einstein angle,

$$\theta_E^2 = \frac{4GM_\bullet}{c^2} \frac{D_{LS}}{D_{OS} D_{OL}} \tag{9.17}$$

where D_{OL} is the observer–lens distance, D_{LS} is the lens–source distance, and D_{OS} is the observer–source distance[3].

The relation between the angular position of the source relative to the observer–lens axis (the optical axis) can be derived from the geometry of the light paths,

$$y = x_{1,2} - 1/x_{1,2} \tag{9.18}$$

where $x_{1,2}$ and y are measured in terms of θ_E and where $x_2 < 0$ by definition. Gravitational lensing conserves surface brightness, and so the magnifications $A_{1,2}$ in the flux of each image relative to that of the unlensed source are proportional to change in the angular area of the source,

$$A_{1,2} = \left| \frac{\partial \vec{y}}{\partial \vec{x}_{1,2}} \right|^{-1} = |1 - x_{1,2}^{-4}|^{-1}. \tag{9.19}$$

The primary image at x_1 is always magnified. The secondary image at x_2 can be demagnified to zero. The two magnifications obey the relations

$$A_1 = A_2 + 1 \geq 1 \tag{9.20}$$

[3] In flat spacetime, which is relevant for Galactic lensing, $D_{OS} = D_{OL} + D_{LS}$. In curved spacetime, which is relevant for cosmological lensing, the distances are the angular diameter distances, and this simple sum no longer holds.
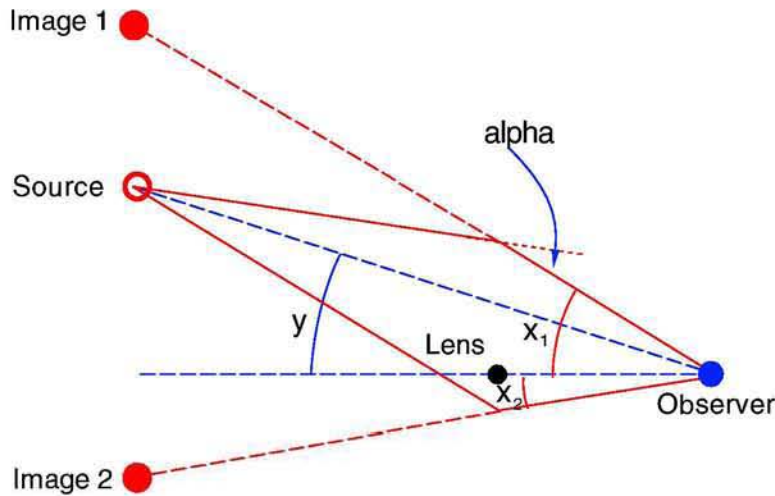
Figure 9.8. The light ray diagram for lensing by a point mass. The light rays from a source, at angular position y relative to the observer–lens optical axis, are bent by the gravitational lens and reach the observer from angles x_1 and x_2, thereby appearing as two images. Unlike a spherical glass lens, the light bending angle α of a gravitational lens is inversely proportional to the impact parameter to the lens (equation (9.16)).

and

$$A \equiv A_1 + A_2 = \frac{y^2 + 2}{y\sqrt{y^2 + 4}}. \tag{9.21}$$

When $y = 0$ the magnification formally diverges and the image appears as a ring of angular size θ_E, the Einstein ring. This divergence is avoided in practice by the finite size of the source (e.g. a star). Finite-sized sources are also sheared tangentially around the Einstein ring as the magnification increases. In the limit of high magnification or small source angle,

$$A \sim 1/y \qquad (y \ll 1). \tag{9.22}$$

9.4.2 Pinpointing the MBH with lensed images

Determining the exact position of the MBH on the IR grid is important because the radio source Sgr A*, which is associated with the MBH, has been detected to date only in one other band, the X-ray (Baganoff *et al* 2001). Currently, the IR position of the radio source Sgr A* is derived indirectly by aligning the radio and IR maps using four maser giants in the inner 15″, which are observed in both bands (Menten *et al* 1997). The exact IR position of the MBH is required, for example, for measuring the IR flux from the MBH, in order to constrain accretion

models; for solving the stellar orbits around the MBH, in order to measure $M_\bullet$ and R_0 (Jaroszyński 1999, Salim and Gould 1999) and to search for general relativistic effects (Jaroszyński 1998, Fragile and Mathews 2000; Rubilar and Eckart 2001); and for detecting the fluctuations of the MBH away from the dynamical center of the stellar cluster, in order to study the stellar potential. Recent measurements of the acceleration vectors of three stars very near Sgr A* provide another way of locating the MBH (Ghez *et al* 2000). The IR/radio alignment and the center of acceleration are close, but do not overlap (figure 9.10), and neither coincide with an IR source. Gravitational lensing can provide a third, independent method for locating the MBH.

When the source, lens and observer move relative to each other, the positions, velocities and magnifications of the images will change with time (figure 9.9). In addition to the requirement that the two images and the lens lie on one line, equations (9.18) and (9.19) imply that the measured angular positions of the two images $\theta_{1,2}$, their projected transverse velocities $v_{t1,2}$ and radial velocities $v_{r1,2}$ relative to the lens, and their measured fluxes $F_{1,2}$, should obey the simple relation

$$-\theta_1/\theta_2 = v_{t1}/v_{t2} = -v_{r1}/v_{r2} = \sqrt{F_1/F_2}. \qquad (9.23)$$

The constraints are based solely on observables, and so are independent of any assumptions about $M_\bullet$, R_0 or the properties of the lensed background sources. The use of equation (9.23) does require knowledge of the exact position of the MBH relative to the stars, since this is needed for measuring the angular distances and for decomposing the radial and tangential components of the velocity. If the MBH position is known, equation (9.23) can be used to search in astrometric measurements of positions, fluxes and velocities for pairs of lensed images around the MBH. Equation (9.23) can also be used to find the position of the MBH on the IR grid, since the MBH lies on the line connecting the two images, and so the intersection of these lines pinpoints its position. This can be done statistically, by enumerating over a grid of trial positions for the MBH, and choosing as the most likely one that which maximizes the number of lensed image pairs.

Figure 9.10 shows the results from such a joint statistical search for the MBH and for a signature of lensing (Alexander 2001). The most likely position of the MBH coincides with the center of acceleration. The random probability for such a likelihood extremum is 0.01. The random probability for such an extremum to fall in either the 1σ error range of IR/radio alignment or that of the center of acceleration is 5×10^{-4}.

The search for the MBH yields also a list of candidate lensed image pairs. The definitive test of lensing is to compare their spectra, which should be identical up to differences due to non-uniform extinction. Unfortunately, spectra for the fainter secondary images are unavailable at this time. Once $M_\bullet$, R_0 and the dust distribution in the Galaxy are assumed, it is possible to derive, albeit with very large uncertainty, the luminosity and distance of the candidate sources. The sources of the two most likely lensed image pair candidates are luminous

$t = -116$ $t = -11$

$t = 0$ $t = +14$
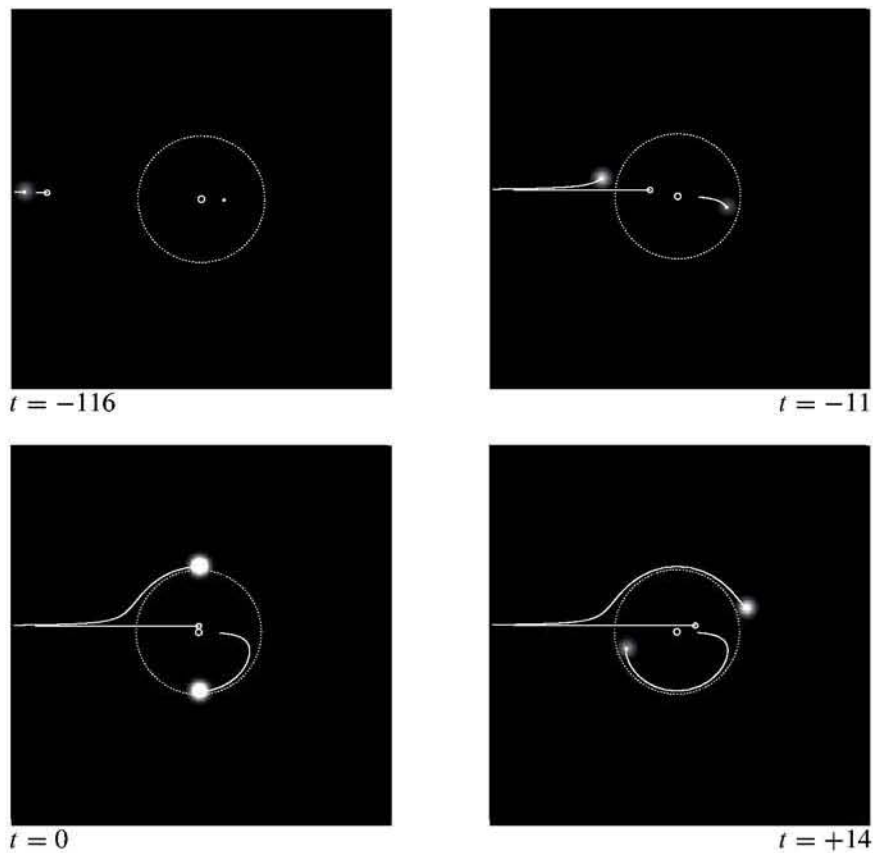
Figure 9.9. A sequence of snapshots simulating the observation of lensing of a background point source by the MBH. Time is in arbitrary units. The background source, which is not observed directly (open circle with straight line tracking the source trajectory) moves in projection from left to right behind the MBH (open circle at center) with an impact parameter of $0.1\theta_E$. The two images (light points with curved lines tracking the image trajectories) move in tandem clockwise about the Einstein ring (large dotted circle). The strongly magnified image (top) is always outside the Einstein ring and is always brighter than the source. The weakly magnified image (bottom) is always inside the Einstein ring and can be strongly demagnified (panels a, b). At peak magnification (panel c) the two images are of comparable brightness (equation (9.20)).

supergiants, a blue supergiant a few kpc behind the Galactic Center and a red supergiant at the far edge of the Galaxy.

This statistical result, while intriguing, requires additional confirmation. Simple models of the distribution of light and dust in the Galaxy predict that the chances of finding luminous supergiants right behind the MBH are very small,
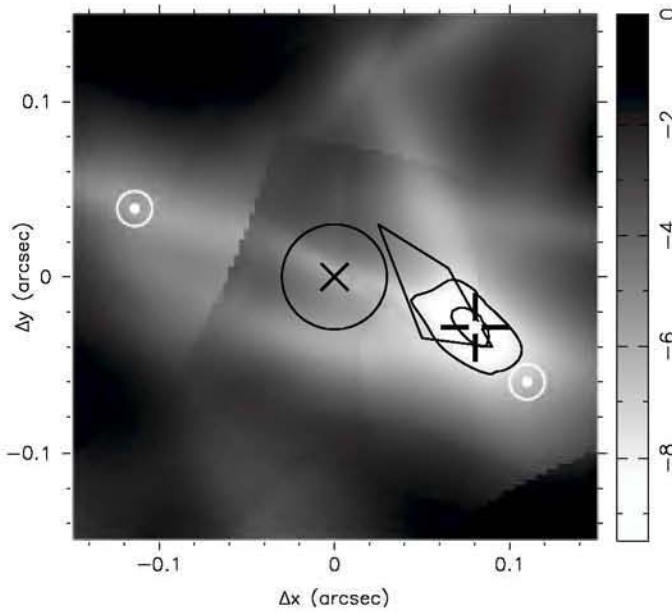
Figure 9.10. Pinpointing the MBH on the IR grid with gravitationally lensed stars (Alexander 2001). A gray scale plot of the logarithm of the likelihood of the MBH position (shifted to 0 at the maximum), for 116 stars from the astrometric compilation by Genzel *et al* (2000), as a function of the shift in the astrometric grid over the central $0.3'' \times 0.3''$ search field. The cross in the center is the origin according to the IR/radio alignment with its 1σ error circle (Menten *et al* 1997). The polygon is the $\sim 1\sigma$ error region for the center of acceleration (Ghez *et al* 2000). The circles are the observed IR sources with their 10 mas error circles. The most likely position of the MBH is indicated by a plus sign with 1σ and 2σ confidence level contours. (Reprinted with permission from *The Astrophysical Journal*.)

and the statistical analysis depends sensitively on the quality of the data and its error properties. Whether or not this particular result survives further scrutiny, it illustrates the potential of gravitational lensing as a tool for the study of the Galactic Center. This statistical method for locating the MBH by gravitational lensing should be re-applied whenever deeper astrometric data become available.

9.4.3 The detection of gravitational lensing

The mode of detection of lensing events depends on the telescope's angular resolution and its photometric sensitivity. When the two images can be resolved, as in the case discussed in section 9.4.2, the phenomenon is called a 'macrolensing' event. When the two images cannot be resolved, only the variability in the flux of the lensed source is observed. This is called a

'microlensing' event. Since θ_E increases with source distance behind the lens, there is a maximal source distance for microlensing, D_μ, which can be estimated by noting that the angular distance between the two images close to peak magnification is $\sim 2\theta_E$, and so

$$D_\mu = \frac{D_{OL}}{(\theta_\infty/\phi)^2 - 1} \tag{9.24}$$

where ϕ is the telescope's angular resolution, $\theta_\infty \equiv \sqrt{4GM_\bullet/c^2 D_{OL}} \sim 1.75''$ is the Einstein angle for a source at infinity, and it is assumed that $\phi < \theta_E$ is the criterion for resolving the two images.

The light curve for a constant velocity trajectory of a background source in the plane of the sky is given by substituting $y(t)$ in equation (9.21),

$$y^2(t) = y_0^2 + \mu^2(t - t_0)^2 \tag{9.25}$$

where y_0 is the impact parameter of the source trajectory relative to the lens, μ is the apparent motion, in units of θ_E per time, and t_0 is the time when $y = y_0$ (figure 9.11). The resulting light curve is symmetric about t_0, and achromatic (i.e. has the same shape in every wavelength). If the photometric sensitivity is large enough to detect the unlensed source, the event will appear as a flaring up of a persistent source; otherwise, it will appear as a transient flare.

In order to plan the observational strategy for detecting gravitational lensing, or to estimate how likely it is that an observed flare is due to lensing, it is necessary to calculate the detection probability. Two quantities are commonly used to express this probability, the optical depth and the lensing rate. The optical depth for gravitational lensing, τ, is usually defined in relation to the probability of having at least one lens along the line of sight. If the cross section of the lens at position z_i is $S(z_i)$, and the number density of lenses there is $n(z_i)$, then the probability P of having at least one lens along the line of sight is the complement of the probability of not encountering any lens,

$$P = 1 - \prod_i (1 - n(z_i)S(z_i)\Delta z_i)$$
$$= 1 - \exp\left(-\int_0^z nS\,dz'\right)$$
$$\equiv 1 - e^{-\tau} \to \tau \qquad (\tau \ll 1). \tag{9.26}$$

It should be emphasized that τ is *not* a probability, and that P and τ are interchangeable only when both are small. It is customary to define the lensing cross section as $S = \pi\theta_E^2$, which corresponds to the region where a source will be magnified by $A > 1.34$ (equation (9.20)). This definition is useful when there are many possible lines of sight and it describes the probability that at *any given instant* a given line of sight will be lensed. This is relevant for Galactic microlensing searches, where millions of background stars are monitored
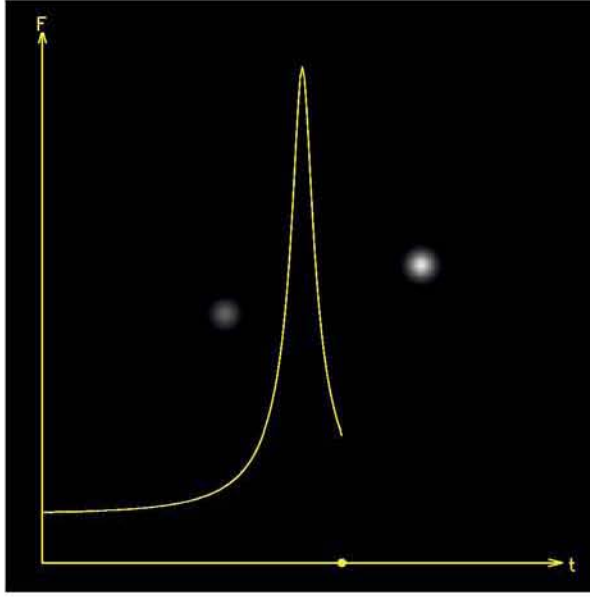
Figure 9.11. The microlensing light curve that corresponds to the lensing event in figure 9.9 up to time $t = 19$, in the case where the two images can not be resolved. The light curve is overlaid on the two lensed images as they would be observed if they could be resolved. The point on the time axis indicates the position of the source along its trajectory. The flux level on the left is the unmagnified flux of the source. The complete light curve will be symmetric relative to the peak flux.

simultaneously to find the rare one that is lensed by an intervening star. The observational situation for gravitational lensing by the MBH in the Galactic Center is different because the position of the lens is known and fixed, and so there is only one line of sight. In analogy to equation (9.26), the optical depth is then defined in relation to the probability of having at least one source behind the lens along the line of sight,

$$\tau = \int_{D_{\text{OL}}}^{\infty} n_{\star} \pi R_{\text{E}}^2 d D_{\text{OS}} \tag{9.27}$$

where R_{E} is the physical size of θ_{E} at the source plane,

$$R_{\text{E}} = \theta_{\infty} \sqrt{(D_{\text{OS}} - D_{\text{OL}}) D_{\text{OS}}}. \tag{9.28}$$

Rough estimates predict $\tau \sim 1$ for lensing by the MBH (assuming no limits on the photometric sensitivity).

The optical depth does not take into account the relative motions of the lens and source, which reshuffle their random alignment and introduces a timescale to

the problem. A more useful quantity for the lensing by the MBH is the lensing event rate with flux above a detection threshold F_0 due to the motion of sources behind the MBH,

$$\Gamma(>F_0) \simeq 2 \int_{D_{OL}}^{\infty} n_* v \frac{R_E}{A} dD_{OS} \qquad A \geq \frac{F_0}{L_*/4\pi D_{OS}^2} \qquad (9.29)$$

where v_* is the source star's projected velocity, L_* is its luminosity and $A \gg 1$ is assumed. For practical applications, equation (9.29) has to be modified to take into account the range of stellar luminosities and velocities, dust extinction, the total duration of the observations T and the sampling rate ΔT (the mean duration of events magnified by more than A is $\bar{t} = \pi R_E/2Av$; only events with $\Delta T < t < T$ can be detected).

Figure 9.12 summarizes the dependence of the lensing cross section, timescale and magnification on D_{LS}. The observational limitations, F_0, T and ΔT, place restrictions on D_{LS} and the impact parameter for which sources can be detected, and affect the typical timescales and peak magnification that are likely to be observed. For example, high magnification events typically have longer timescales because the source trajectory must have a smaller impact parameter and so it spends more time in the Einstein radius. Therefore, observations with limited temporal sampling will tend to pick out high magnification events.

Were any microlensing events detected? A couple of possible transient flaring events were detected very close to Sgr A* (Genzel *et al* 1997, Ghez *et al* 1998). For one of these a light curve was recorded, but as it was under-sampled only estimates of a timescale ($\sim$1 yr) and a typical magnification ($\sim A > 5$) could be derived from it. The *a posteriori* probability of detecting a lensing event was estimated at only 0.5%, but on the other hand, the observed timescale and magnification are close to the median value that is expected for the observational limitations (Alexander and Sternberg 1999). The interpretation of this event remains inconclusive.

9.4.4 Magnification bias

A lens magnifies by enlarging the angular size of the unlensed sky behind it, and since surface brightness is conserved, the fluxes of sources are magnified by the same amount. When the photometric sensitivity is such that all the stars can be detected even without being magnified, then the effect of lensing is to decrease the surface density of sources. However, if the fainter stars cannot be observed unless magnified, there are two possibilities (figure 9.13): either there are enough faint sources that are magnified above the detection threshold to over-compensate for the decrease in surface density ('positive magnification bias'), or there are not enough faint sources ('negative magnification bias'). The lensed luminosity function (number of stars per flux interval) is related to the unlensed one by

$$\left. \left(\frac{d\Sigma}{dF} \right)_{\text{lensed}} \right|_F = A^{-2} \left. \left(\frac{d\Sigma}{dF} \right)_{\text{unlensed}} \right|_{F/A} \qquad (9.30)$$
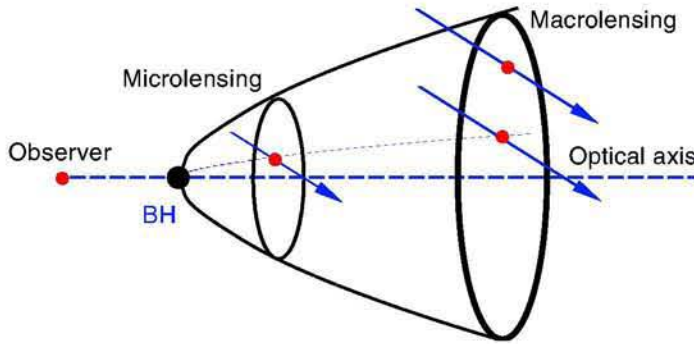
Figure 9.12. The lensing cross section, timescale and magnification as a function of source distance behind the MBH. The size of the Einstein ring, or lensing cross section (cone, see equation (9.28)), increases with distance behind the MBH. Close behind the MBH the Einstein ring (and the distance between the two images) is smaller than the telescope's angular resolution and the lensing appears as a microlensing event. Farther out, the Einstein ring is large enough for the two images to be resolved, and the lensing appears as a macrolensing event. The duration and peak magnification of the events depend on the impact parameter of the stellar trajectories (arrows). The closer they are to the optical axis, the longer the events and the higher the peak magnification are. Trajectories with impact parameters at a fixed ratio of the Einstein radius (the two trajectories connected by the dotted line) will have the same peak magnification (equation (9.21)), but the event duration will be longer for the sources farther away behind the MBH (assuming a uniform velocity field).

where Σ is the surface number density of stars and F the flux. In many cases the luminosity function is well approximated by a power law, $d\Sigma/dF \propto F^{-\beta}$. It then follows from equation (9.30) that for $\beta = 2$ the decrease in the total surface density is exactly balanced by the magnification of faint stars above the detection threshold.

The chances for the detection of this effect in the Galactic Center appear small. A statistically meaningful detection requires a very high surface density that probably exceeds even that around the MBH (Wardle and Yusef-Zadeh 1992), and furthermore, models of the stellar luminosity function in the inner Galactic Center suggests that $\beta \sim 2$ for giants (Alexander and Sternberg 1999).

9.4.5 Beyond the point mass lens approximation

Up to now, we have considered only the simple case of lensing by a point mass. There are two reasons to explore more complicated models. The first is that it

Source plane   Image plane
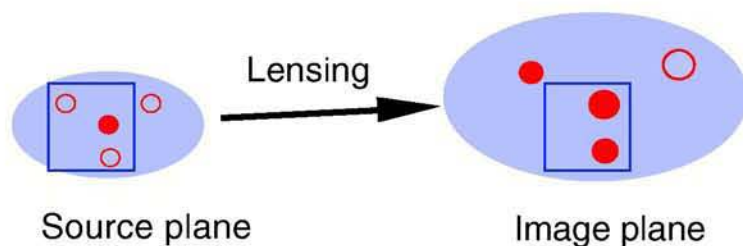
Figure 9.13. The magnification bias in star counts due to gravitational lensing. Stars (circles) are counted in a field of a fixed angular area (square) with a telescope of a given photometric detection threshold. The field in the unlensed sky (source plane, left), contains one bright star that can be observed (filled circle) and two stars that are too faint to be observed (open circles). Gravitational lensing stretches angular areas and magnifies fluxes by the same factor. The field in the lensed sky (image plane, right), now contains only two stars, but because they are magnified, both can be observed. This is an example of positive magnification bias, where gravitational lensing increases the *apparent* stellar surface density even as it decreases the *total* surface density. Negative magnification bias occurs when there are not enough faint stars in the lensed population to compensate for the decrease in the total surface density.

would be useful if gravitational lensing could be used to dispel any remaining doubts that the dark compact mass in the Galactic Center is indeed a MBH, and not some other extended distribution of matter, such as a compact cluster of stellar remnants (Maoz 1998) or a concentration of exotic particles (Tsiklauri and Viollier 1998). Unfortunately, it can be shown that the behavior of high-magnification light curves near peak magnification is universal and independent of the details of the lens (equation (11.21b) of Schneider *et al* 1992). For spherically symmetric mass distributions this implies that the light curves differ only in the low magnification tails, which are much harder to observe. The second reason is that the MBH is surrounded by a massive stellar cluster. Because the stellar mass is not smoothly distributed but is composed of discrete point masses, its effect on the lensing properties of the MBH is much larger than one may naively estimate by adding the stellar mass to that of the MBH. We conclude the discussion of gravitational lensing in the Galactic Center by describing briefly the effect of enhanced lensing by stars near the MBH (Alexander and Loeb 2001, Chanamé *et al* 2001).

The effect of stars on lensing by the MBH is similar to that of planets on microlensing by a star, an issue that was studied extensively for the purpose of detecting planets by microlensing (e.g. Gould and Loeb 1992). The lensing cross section of an isolated star is $\theta_E^2(M_\star)/\theta_E^2(M_\bullet) = M_\star/M_\bullet \lesssim 10^{-6}$ smaller than that of the MBH (equation (9.17)). However, when the star lies near $\theta(M_\bullet)$, the shear of the MBH distorts its lensing cross section, which develops
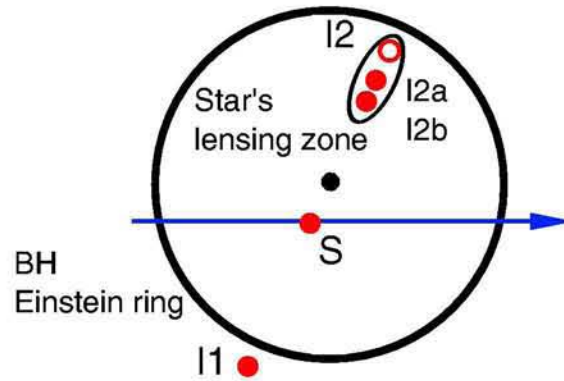
Figure 9.14. A schematic representation of lensing enhancement by a star near the MBH (Alexander and Loeb 2001, Chanamé *et al* 2001). A light source S passes behind the MBH (filled circle in the middle) and, in the absence of any other lensing mass, appears as two images: I1 outside the Einstein ring and I2 inside the Einstein ring. When one of the stars near the MBH happens to lie (in projection) close to I2, it will split I2 into two or four sub-images (two shown here), I2a and I2b. The star's Einstein ring is sheared by the potential of the MBH to an elongated shape of complex topology (represented here for simplicity as an ellipse), which increases in size the nearer I2 and the star are to the Einstein radius of the MBH. This effect increases the cross section for high-magnification events above that of an isolated MBH, and changes the character of the light curves.

a complex topology, becomes radially elongated and is increased by up to an order of magnitude (figure 9.14). As the stars orbit the MBH, their elongated cross sections scan the lens plane. If these happen to intersect one of the images of a background source that is lensed by the MBH, the image will be split into two or four sub-images whose angular separation will be of order $\theta_E(M_\star)$, and so the sub-images will not be individually resolved. However, their combined flux will be significantly magnified. This will increase the probability of high-magnification events over what is expected for lensing by the MBH alone. The light curves of such events will no longer be symmetric as they are for a point mass, but will exhibit a complex structure (e.g. Wambsganss 1997), and their typical variability timescales will rise sharply for images that lie near $\theta_E(M_\bullet)$ because of the increased stellar cross section for lensing. Enhanced lensing by stars in the Galactic Center is estimated to increase the probability of $A > 5$ lensing events by ~ 2 and of $A > 50$ events by ~ 3.

9.5   Summary

Observations of the MBH in the Galactic Center present a unique opportunity to study the consequences of extreme stellar density, velocity and tidal fields on the dynamics and evolution of stars and their relation to the dynamics and evolution of the MBH. The existence of a high-density relaxed stellar cusp around the MBH in the Galactic Center is theoretically motivated, and supported by observations. We explored some of the consequences of this environment for the appearance, internal structure and evolution of stars, through exotic object formation by direct collisions, collisional destruction of giant envelopes, stochastic tidal spin-up of stars by collisions with other stars, and extreme tidal interactions in the course of tidal scattering by the MBH. It was shown that tidal processes have the potential of affecting a significant fraction of the stars over a large volume around the MBH.

The MBH is also a gravitational lens. This can be used to probe the dark mass and the stars around it, but it also has the potential for complicating the interpretation of observations in the Galactic Center. Different detection modes were considered: macrolensing, microlensing, magnification bias, and the detection probability and detection rate were defined. Results from a statistical method for detecting lensed images and for pinpointing the MBH on the IR grid suggest that there may be a few far background supergiants that are lensed by the MBH. We described a lensing effect that involves both the MBH and the stars around it, and can increase the probability of high-magnification events and modify the structure of the light curves.

The topics covered by this chapter by no means exhaust the scope of the subject. We did not address, among others, star formation near the MBH, the role of stellar evolution in feeding the MBH, or compact stellar remnants and x-ray sources. Some of these issues are discussed elsewhere in this book.

Over the next decade a wide array of IR instruments, both ground based and space borne, will improve the quality of photometric, spectroscopic and astrometric observations of the Galactic Center by orders of magnitude. Many of the issues discussed here will be resolved, as new questions will surely be raised. One thing is certain—we can look forward to exciting times in Galactic Center research.

References

Alexander T 1999 *Astrophys. J.* **527** 835
——2001 *Astrophys. J.* **553** L149
Alexander T and Kumar P 2001 *Astrophys. J.* **549** 948
Alexander T and Livio M 2001 *Astrophys. J.* **560** L143
Alexander T and Loeb A 2001 *Astrophys. J.* **551** 223
Alexander T and Sternberg A 1999 *Astrophys. J.* **520** 137
Ayal S, Livio M and Piran T 2000 *Astrophys. J.* **545** 772
Baganoff F K *et al* 2001 *Nature* **413** 45

Bahcall J N and Wolf R A 1976 *Astrophys. J.* **209** 214

——1977 *Astrophys. J.* **216** 883

Binney J and Tremaine S 1987 *Galactic Dynamics* (Princeton, NJ: Princeton University Press) p 520

Blum R D, Sellgren K and DePoy D L 1996 *Astrophys. J.* **470** 864

Chanamé J, Gould A and Miralda-Escudé J 2001 *Astrophys. J.* **563** 793

Eckart A and Genzel R 1997 *Mon. Not. R. Astron. Soc.* **284** 576

Fragile P C and Mathews G J 2000 *Astrophys. J.* **542** 328

Frank J 1978 *Mon. Not. R. Astron. Soc.* **184** 87

Frank J and Rees M J 1976 *Mon. Not. R. Astron. Soc.* **176** 633

Freitag M and Benz W 2002 *Astron. Astrophys.* **394** 345

Gebhardt K, Rich R M and Ho L C 2002 *Astrophys. J.* **578** L41

Genzel R, Eckart A, Ott T and Eisenhauer F 1997 *Mon. Not. R. Astron. Soc.* **291** 219

Genzel R, Hollenbach D and Townes C H 1994 *Rep. Prog. Phys.* **57** 417

Genzel R, Pichon C, Eckart A, Gerhard O E and Ott T 2000 *Mon. Not. R. Astron. Soc.* **317** 348

Genzel R, Thatte N, Krabbe A Kroker H and Tacconi-Garman L E 1996 *Astrophys. J.* **472** 153

Ghez A M, Klein B L, Morris M and Becklin E E 1998 *Astrophys. J.* **509** 678

Ghez A M, Morris M, Becklin E E, Tanner A and Kremenek T 2000 *Nature* **407** 349

Gould A and Loeb A 1992 *Astrophys. J.* **396** 104

Hills J G 1975 *Nature* **254** 295

Jaroszyński M 1998 *Acta Astron.* **48** 653

——1999 *Astrophys. J.* **521** 591

Komossa S and Bade N 1999 *Astron. Astrophys.* **343** 775

Komossa S and Greiner J 1999 *Astron. Astrophys.* **349** 45

Lauer T R, Faber S M, Ajhar E A, Grillmair C J and Scowen P A 1998 *Astrophys. J.* **116** 2263

Lightman A P and Shapiro S L 1977 *Astrophys. J.* **211** 244

Magorrian J and Tremaine S 1999 *Mon. Not. R. Astron. Soc.* **309** 447

Maoz E 1998 *Astrophys. J.* **494** L181

Menten K M, Reid M J, Eckart A and Genzel R 1997 *Astrophys. J.* **475** L111

Monaghan J J 1992 *Annu. Rev. Astron. Astrophys.* **30** 543

Morris M 1993 *Astrophys. J.* **408** 496

Murphy B W, Cohn H N and Durisen R H 1991 *Astrophys. J.* **370** 60

Press W H and Teukolsky S A 1977 *Astrophys. J.* **213** 183

Reid M 1993 *Annu. Rev. Astron. Astrophys.* **31** 345

Renzini A, Greggio L, di Serego-Alighieri S, Cappellari M, Burstein D and Bertola F 1995 *Nature* **378** 39

Rubilar G F and Eckart A 2001 *Astron. Astrophys.* **374** 95

Salim S and Gould A 1999 *Astrophys. J.* **523** 633

Schneider P, Ehlers J and Falco E E 1992 *Gravitational Lenses* (New York: Springer)

Thorne K S and Zytkow A N 1975 *Astrophys. J.* **199** 19

Tsiklauri D and Viollier R D 1998 *Astrophys. J.* **500** 591

Ulmer A, Paczyński B and Goodman J 1998 *Astron. Astrophys.* **333** 379

Wambsganss J 1997 *Mon. Not. R. Astron. Soc.* **284** 172

Wardle M and Yusef-Zadeh F 1992 *Astrophys. J.* **387** L65

Young P 1980 *Astrophys. J.* **242** 1232

Chapter 10

Black hole accretion models for the Galactic Center

Robert F Coker
Los Alamos National Laboratory, USA

Now we turn our attention to the hydrodynamic process of gas falling onto weakly active black holes and the associated emission. Spherical accretion from the surrounding gas (Bondi–Hoyle accretion) as well as rotating and radiatively inefficient accretion (Advection-Dominated Accretion Flows, ADAFs) are discussed. The basic hydrodynamical equations for these problems are derived and applied to the Galactic Center.

10.1 Introduction

Compact objects gravitationally capture matter in a process known as accretion. Matter falling down the steep gravitational potential of a compact object may release more than 10% of its rest-mass as radiation. In fact, matter accreting via a disk onto a maximally spinning black hole can release up to 42% of its rest mass! By way of comparison, nuclear burning which converts H to Fe releases a maximum of 0.9%. Thus, accretion can be an extremely efficient energy source. In general, the accretion geometry is determined by the interplay between the heating and cooling mechanisms, the intrinsic angular momentum present in the accreting gas, and the inner and outer boundary conditions of the flow. However, it is not easy to generically determine the accretion geometry onto a compact object and then calculate the resulting emitted spectrum. This requires solving the time-dependent three-dimensional relativistic MHD (magnetohydrodynamic) equations with radiative transfer, a formidable task indeed. Therefore, we will focus on idealized accretion models which use assumptions to tremendously simplify the overall problem. In addition, in order to apply our discussion to

276

Sgr A*, the supposed supermassive black hole in the Galactic Center, we deal here only with black hole accretion.

The models discussed here are fairly general but in practice are mainly applicable to black holes at relatively low (sub-Eddington) accretion rates. Standard thin disk accretion models, for example, which are invoked for luminous quasars are not explicitly discussed as they can be found in many textbooks (e.g. Frank *et al* 2002).

10.2 Accreting gas with zero angular momentum

10.2.1 Adiabatic spherical accretion

The simplest black hole accretion scenario was worked out more than half a century ago (Hoyle and Lyttleton 1939, Bondi and Hoyle 1944, Bondi 1952). First, we assume that the mean free path of the accreting gas is small so that the flow is hydrodynamical; that is, some mechanism, such as plasma instabilities, serves to effectively couple the gas particles. Second, we assume that the gravitational field of the black hole is dominant so that the field is spherically symmetric and the self-gravity of the accreting gas can be ignored. Third, we assume the black hole has zero charge and is non-rotating. Fourth, we assume the accreting gas has zero angular momentum and is at rest at infinity. Fifth, we ignore magnetic fields and entropy loss due to radiation so that the accreting gas can be approximated as adiabatic. These assumptions allow us to apply solutions of the steady-state non-relativistic hydrodynamic equations.

If the accreting particles are assumed to be collisionless, the gas is no longer a hydrodynamical fluid. As a result, compared to the hydrodynamical models discussed in this chapter, the density near the black hole and thus the mass accretion rate will be depressed by a factor of $(c_\infty/c)^2$. Collisionless accretion thus results in a density profile that goes as $r^{-1/2}$ rather than $r^{-3/2}$. A flatter density profile leads to considerably lower mass accretion rates. Since the primary problem in explaining the spectrum of Sgr A* is its low luminosity for its accretion rate as derived from hydrodynamical models, perhaps non-hydrodynamical models which assume larger particle mean-free paths may prove useful. Of course, then the problem becomes one of explaining why the accretion flow around Sgr A* is collisionless while those of other accreting black hole systems are apparently not.

The mass continuity equation for spherical hydrodynamical accretion, given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0 \qquad (10.1)$$

where ρ is the mass density of the accreting gas and $\boldsymbol{u}$ is the radial velocity

(denoted so that for inward flow, $u > 0$), reduces to

$$\frac{1}{r^2}\frac{d}{dr}(r^2 \rho u) = 0 \tag{10.2}$$

or

$$\frac{2}{r} + \frac{\rho'}{\rho} + \frac{u'}{u} = 0 \tag{10.3}$$

where the prime ($'$) denotes differentiation with respect to r (d/dr).

The non-relativistic momentum or Euler equation is given by

$$\frac{du}{dt} + \frac{1}{\rho}\nabla P + \nabla \Phi = 0 \tag{10.4}$$

where P is the thermal pressure and Φ is the gravitational potential. The total or Lagrangian time derivative is given by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u \cdot \nabla. \tag{10.5}$$

Since it is assumed that

$$P \propto \rho^\gamma \tag{10.6}$$

where γ is the adiabatic index, the adiabatic sound speed, c_s, is determined from

$$c_s^2 \equiv \frac{dP}{d\rho} = \frac{dP}{dr}\left(\frac{d\rho}{dr}\right)^{-1} \tag{10.7}$$

or $c_s = \sqrt{\gamma P/\rho}$. In terms of the temperature, T, of a fully ionized plasma consisting of pure hydrogen, the adiabatic sound speed can be written as

$$c_s^2 = \frac{2\gamma k_B T}{m_p} \tag{10.8}$$

where k_B is the Boltzmann constant and m_p is the mass of the proton. For a point mass of mass M, we have

$$\nabla \Phi = \frac{GM}{r^2} \tag{10.9}$$

where G is the gravitational constant. Assuming a steady state, the Euler equation (10.4) then becomes

$$u\frac{du}{dr} + \frac{1}{\rho}\frac{dP}{dr} + \frac{GM}{r^2} = 0 \tag{10.10}$$

or

$$uu' + c_s^2\frac{\rho'}{\rho} + \frac{GM}{r^2} = 0. \tag{10.11}$$

Solving for ρ' and u' using (10.3) and (10.11), we find

$$u' = u \left(\frac{2c_s^2}{r} - \frac{GM}{r^2} \right) (u^2 - c_s^2)^{-1} \tag{10.12}$$

and

$$\rho' = \rho \left(\frac{GM}{r^2} - \frac{2u^2}{r} \right) (u^2 - c_s^2)^{-1}. \tag{10.13}$$

A relativistic solution to the equations of motion (Shapiro and Teukolsky 1983, appendix G) shows that

$$u \to \sqrt{\frac{2GM}{r}} \equiv u_{\mathrm{ff}} \tag{10.14}$$

as $r \to 0$; that is, moving inward, u must rise monotonically from 0 to the free-fall velocity. Given this relation, in order to avoid singularities in the flow, there must exist a transonic radius r_t such that the numerators and denominators in (10.12) and (10.13) simultaneously vanish. This occurs when

$$\frac{GM}{2r_t} = u^2(r_t) = c_s^2(r_t). \tag{10.15}$$

Thus, at the transonic radius, the internal thermal energy per unit mass of the accreting gas is comparable to the gravitational potential. In order to evaluate (10.15) in terms of the boundary values at infinity, we integrate (10.11) using (10.7) and find the Bernoulli equation:

$$\frac{u^2}{2} + \frac{c_s^2}{\gamma - 1} - \frac{GM}{r} = \frac{c_\infty^2}{\gamma - 1} \tag{10.16}$$

where c_∞ is c_s at infinity. Note that (10.16) also shows that the velocity approaches the free-fall velocity near the origin. In fact, at the Schwartzchild radius $r_s \equiv 2GM/c^2$ where c is the speed of light, the velocity must approach c. Combining (10.15) and (10.16), we find

$$r_t = \frac{5 - 3\gamma}{8} R_A \tag{10.17}$$

where the so-called 'accretion radius' is defined as

$$R_A \equiv \frac{2GM}{c_\infty^2} \tag{10.18}$$

and

$$u(r_t) = c_s(r_t) \equiv c_t = \frac{2}{5 - 3\gamma} c_\infty^2. \tag{10.19}$$

We can now calculate the mass accretion rate, an important diagnostic for accretion models. In this simple picture, the mass continuity equation (10.2) shows that $r^2 \rho u$ is a constant for all r. Integration gives

$$\dot{M} \equiv 4\pi r^2 \rho u. \tag{10.20}$$

Using (10.7) again, one can show that

$$\frac{\rho}{\rho_\infty} = \left(\frac{c_s}{c_\infty}\right)^{2/(\gamma-1)}. \tag{10.21}$$

Combining the last four equations, the mass accretion rate is seen to be

$$\dot{M} = \pi \lambda R_A^2 \rho_\infty c_\infty \tag{10.22}$$

where

$$\lambda \equiv 2^{(9-7\gamma)/(\gamma-1)/2}(5-3\gamma)^{(3\gamma-5)/(\gamma-1)/2}. \tag{10.23}$$

For $\gamma = 5/3$, $\lambda = 0.25$; λ monotonically increases towards a value of $e^2/4 = 1.12$ as $\gamma \to 1$. The definition for the accretion radius R_A is now clear: to within the factor λ, the mass accretion rate is simply the mass flux of gas with density ρ_∞ and velocity c_∞ through the area of a circle with a radius R_A.

We now can construct complete accretion profiles for the velocity, density, and temperature. Far from the black hole, gravity is unimportant so that ρ and T are approximately constant and equal to their values at infinity. From (10.20) and (10.22), the velocity for $r \gg r_t$ is thus

$$u = \frac{c_\infty \lambda}{4}\left(\frac{R_A}{r}\right)^2. \tag{10.24}$$

Once through the transonic point, the gas becomes supersonic and the velocity approaches free fall. Thus, for $r \ll r_t$, using (10.22) and (10.18), the density profile approaches

$$\rho = \frac{\rho_\infty \lambda}{4}\left(\frac{R_A}{r}\right)^{3/2} \tag{10.25}$$

while from (10.21) the temperature profile $(T \propto c_s^2)$ is

$$T = T_\infty \left(\frac{\lambda}{4}\right)^{\gamma-1}\left(\frac{R_A}{r}\right)^{3(\gamma-1)/2}. \tag{10.26}$$

In adiabatic spherical accretion, the only length-scale in the problem is the accretion radius, R_A. Since the mass of the black hole, M, only enters into the problem via R_A, one needs only to specifiy R_A, ρ_∞, and c_∞ to completely determine the solution.
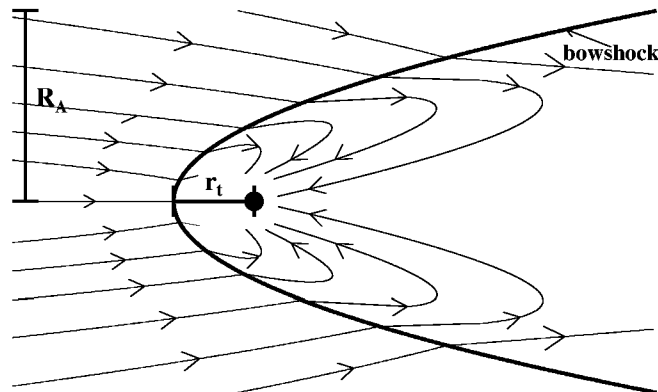
Figure 10.1. Schematic diagram showing streamlines of initially supersonic particles accreting onto a compact object.

Transonic accretion of spherically symmetric and initially subsonic gas is known as Hoyle–Lyttleton accretion. However, if the accreting gas has non-zero velocity at infinity relative to the black hole, these results can still be applied with only slight modification. Such a model can be applicable to a black hole moving rapidly through the ISM or a black hole being fed by stellar winds from a large star cluster located far enough away from the black hole that the flow near the black hole is approximately uniform and planar. If the velocity of the gas relative to the black hole is given by v_∞, then replacing c_∞ with $\sqrt{c_\infty^2 + v_\infty^2}$ in the Hoyle–Lyttleton results will be correct to within a correction factor of order unity (the slight asymmetry between c_∞ and v_∞ can be seen in the Bernoulli equation).

If the flow is initially supersonic so $v_\infty > c_\infty$, a bowshock will form at r_t; at distances closer than $\simeq r_s$, the flow will be nearly spherically symmetric. Figure 10.1 shows a schematic diagram of what a supersonic accretion flow would look like in the rest frame of the black hole. Note that particles on streamlines that are within R_A of the black hole will eventually be accreted. In the hypersonic limit, where $v_\infty \gg c_\infty$, the accretion of gas with zero net angular momentum is known as Bondi–Hoyle accretion.

For Sgr A*, the black hole in the Galactic Center, one idea is that the primary source for the accreting material is stellar winds from young massive stars that are distributed fairly uniformly around the black hole at distances comparable to R_A. Thus, although the flow is probably supersonic and so of Bondi–Hoyle type, the assumption of planarity—and thus probably that of zero net angular momentum—

is likely to be at least somewhat inappropriate.

10.2.1.1 A comment on the relativistic solution

This derivation is only valid in the non-relativistic limit: it has been assumed that $k_B T_\infty \leq k_B T(r_t) \ll m_p c^2$ and $r_t \gg r_s$. From (10.17) and (10.19) it can be seen that as $\gamma \to 5/3$, $r_t \to 0$ and so the derivation breaks down. However, a general relativistic calculation only changes the gas profiles by constants that are on the order of unity; this is expected since the flow is determined by the boundary conditions at infinity, where the gravity of the compact object is negligible. Such a derivation can be done in the same manner as here as long as r is identified with the Schwarzschild radial coordinate, u with the radial four-velocity, and ρ with the rest-mass density. The primary difference, as shown in Shapiro and Teukolsky (1983), is that the transonic point shifts.

Specifically, (10.17) becomes

$$r_t = \frac{1 + 3c_\infty^2}{4} \frac{c_\infty}{c_t} R_A \tag{10.27}$$

and (10.19) becomes

$$(1 + 3c_t^2)(\gamma - 1 - c_t^2)^2 = (\gamma - 1 - c_\infty^2)^2. \tag{10.28}$$

In this section only, for ease of writing, speeds (c_∞ and c_t) are written in units of c. After a lot of algebra, (10.28) can be rewritten as

$$x^3 + a_1 x + a_2 = 0 \tag{10.29}$$

where

$$x \equiv c_t^2 + \frac{7 - 6\gamma}{9} \tag{10.30}$$

$$a_1 \equiv -\frac{1}{27}(3\gamma - 2)^2 \tag{10.31}$$

and

$$a_2 \equiv \frac{2}{81}\left(3\gamma^3 - 6\gamma^2 + 4\gamma - \frac{8}{9}\right) - \frac{(\gamma - 1 - c_\infty^2)^2}{3}. \tag{10.32}$$

If

$$\frac{a_1^3}{27} + \frac{a_2^2}{4} < 0 \tag{10.33}$$

then (10.29) has three real solutions, only one of which is physically applicable here (e.g. one solution corresponds to a complex sound speed). It can be shown that for $\gamma = 1$, if $c_\infty^2 > 2/(9\sqrt{3})$, there is no accretion solution. This limit increases with increasing γ until at $\gamma \simeq 1.5$, there is a solution for all $c_\infty < 1$.
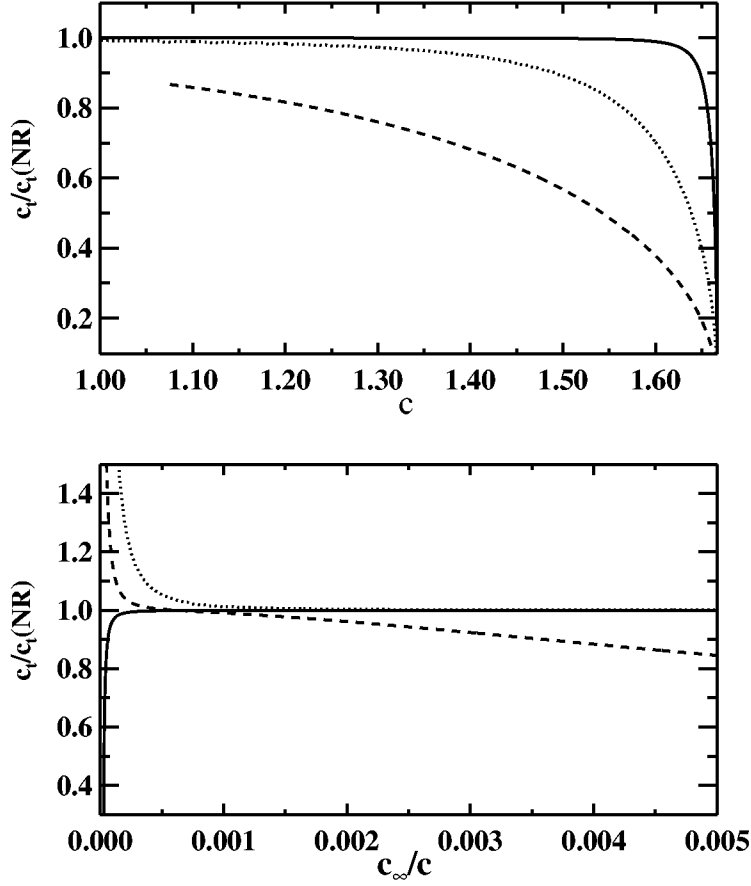
Figure 10.2. Transonic soundspeed determined from (10.28) and normalized to the non-relativistic solution (10.19). The top curves are functions of γ for $c_\infty = 0.01$ (full), 0.1 (dotted), and 0.5 (dashed). The bottom curves are functions of c_∞ for $\gamma = 1.01$ (full), 1.33 (dotted), and 1.66 (dashed). Note the latter curves are for small values of c_∞.

If an accretion solution for c_t exists, it is given by

$$\phi \equiv \arccos\left(-\frac{a_2}{2}\left(\frac{-a_1^3}{27}\right)^{-1/2}\right) \tag{10.34}$$

$$s_1 \equiv 2\sqrt{\frac{-a_1}{3}}\cos\left(\frac{\phi}{3}\right) - \frac{7-6\gamma}{9} \tag{10.35}$$

$$s_2 \equiv 2\sqrt{\frac{-a_1}{3}} \cos\left(\frac{\phi}{3} + \frac{4}{3}\pi\right) - \frac{7 - 6\gamma}{9}. \tag{10.36}$$

If $\gamma - c_\infty^2 < 1$, then $c_t = \sqrt{s_1}$; otherwise, $c_t = \sqrt{s_2}$. Figure 10.2 shows some solutions for the transonic sound speed normalized to the non-relativistic solution (10.19). Only very near $\gamma = 5/3$ *and* $c_\infty = 1$ does the solution approach the first order relativistic solution of $\sqrt{c_\infty 2/3}$. Further, substantial deviations from (10.19) are evident for a variety of values of γ (but particularly near $\gamma = 5/3$), even for small values of c_∞. As a result, the location of the bowshock will differ from (10.17) even for non-relativistic sound speeds.

10.2.2   Supersonic non-adiabatic spherical accretion

The Bondi–Hoyle accretion model provides analytical expressions which give a good first look at the accretion process, but it is simplistic and makes many assumptions that are invalid for Sgr A*. For example, the accreting gas near Sgr A* is likely to be relativistic, but the gas at large radius is not. Also, the gas cools via radiation and heats via the reconnection of magnetic field lines. Thus, γ is not constant and an adiabatic solution is inaccurate. Since most of the radiation may be coming from near the black hole, we will set up the relativistic hydrodynamic equations which can then be numerically evaluated. In order to make the problem tractable we assume the gas consists of fully ionized hydrogen so that $n_i = n_e = n$.

We replace (10.4) with the relativistic Euler equation for a spherical geometry:

$$uu' + HP' + \frac{GM}{r^2} = 0 \tag{10.37}$$

where the pressure is given by

$$P = 2nk_BT \tag{10.38}$$

and

$$H \equiv \frac{c^2 + u^2 - 2GM/r}{P + e_\rho + \epsilon}. \tag{10.39}$$

In the non-relativistic limit, H reduces to $1/\rho$. The particle mass-energy density is

$$e_\rho = m_pc^2n \tag{10.40}$$

and the internal energy density of the gas is

$$\epsilon = \alpha nk_BT. \tag{10.41}$$

In the fully ionized but non-relativistic limit, i.e. $10^5 \text{ K} \lesssim T \lesssim 6 \times 10^9 \text{ K}, \alpha = 3$. However, in the relativistic electron limit, but still non-relativistic for the protons,

so that 6×10^9 K $\lesssim T \lesssim 10^{13}$ K, $\alpha = 9/2$. For a smooth transition, we use the general expression from Chandrasekhar (1939) that is valid for all T:

$$\alpha = x \left(\frac{3K_3(x) + K_1(x)}{4K_2(x)} - 1 \right) + y \left(\frac{3K_3(y) + K_1(y)}{4K_2(y)} - 1 \right) \qquad (10.42)$$

where $x \equiv m_e c^2 / k_B T$, $y \equiv m_p c^2 / k_B T$ and K_i refers to the ith order modified Bessel function. If the gas is everywhere non-relativistic and in the absence of heating and cooling, these equations reduce to a $\gamma = 5/3$ adiabatic gas.

To include the braking effects of radiation pressure, we include a term $P_{\rm rad}$ in the pressure (10.38) and a term $3P_{\rm rad}$ in the internal energy density (10.41) where

$$P_{\rm rad} = \int_0^{\nu_m} \frac{4\pi}{3} u_P \, d\nu. \qquad (10.43)$$

Here, ν_m is the frequency below which the radiative emission at that radius is highly absorbed. Thus, the optical depth, $\tau_r^{\infty}(\nu_m)$, from r to infinity is approximately unity. The Planck energy distribution, u_P, in the low-frequency Rayleigh–Jeans limit is simply $2\nu^2 k_B T / c^3$, so that

$$P_{\rm rad} = \frac{8\pi}{9} k_B T \left(\frac{\nu_m}{c} \right)^3. \qquad (10.44)$$

In the Rayleigh–Jeans limit, $h\nu_m \ll k_B T$; for Sgr A* we find that this is always true. In general, ν_m reaches a maximum value of $\lesssim 10^{12}$ Hz near the black hole but drops off roughly as r^{-1} so that as long as $n \gtrsim 10^5$ cm^{-3} near the black hole, the effects of radiation pressure are negligible. This is expected since the luminosity of Sgr A* is vastly sub-Eddington (see section 10.2.3).

Substituting the definitions (10.38), (10.40), and (10.41) into (10.37) and using (10.3), we solve for the derivative of the velocity

$$u' = u \frac{2nk_B(T' - 2T/r)H + GM/r^2}{2nk_B T H - u^2}. \qquad (10.45)$$

For simplicity, we write u' as the sum of two terms:

$$u' = fT' + g \qquad (10.46)$$

where

$$f \equiv \frac{u2nk_B H}{2nk_B T H - u^2} \qquad (10.47)$$

and

$$g \equiv \left(\frac{u}{r} \right) \frac{GM/r - 4nk_B T H}{2nk_B T H - u^2}. \qquad (10.48)$$

The form of these expressions is that of the classic wind equations (e.g. Parker 1960, Melia 1988). In the model we consider here, the gas is supersonic at

infinity and remains supersonic throughout its inward trajectory. We therefore avoid the special handling required for solutions that cross any sonic points, where the denominator of (10.45) vanishes.

With heating and cooling, we can no longer use (10.6) for our equation of state. We use a modified form of the first law of thermodynamics:

$$u\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\epsilon}{n}\right) + uP\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{n}\right) + \frac{\Gamma - \Lambda}{n} = 0 \tag{10.49}$$

where Γ and Λ are the frequency-integrated heating and cooling emissivities, respectively.

These equations assume that the magnetic field is radial; that is, $\boldsymbol{u} \times \boldsymbol{B} = 0$ and there is no large-scale current. Thus, the compression of the gas is parallel to the magnetic field lines and we can ignore the magnetic pressure. If the magnetic field at large radius is assumed to originate from multiple stellar sources, it will be tangled and the three components of the field will be roughly equal. As the gas accretes, it is compressed radially and, in the absence of any dissipation or reconnection, the radial component will grow as $1/r^2$ while the tangential components will only grow as $1/r$. Therefore, for $r \lesssim R_A$ we can assume the magnetic field is predominantly radial; in all that follows we ignore the tangential component of the magnetic field. However, due to the 'no hair' theorem, the magnetic field must be purely tangential at the event horizon of a black hole; we shall assume for now that this occurs infinitesmally close to the event horizon. In this model, the magnetic field serves only to determine Γ, the heating term.

Substituting in (10.3) and the expressions for P and ϵ, dividing by u, taking the derivative, and rearranging gives

$$\alpha k_{\mathrm{B}} T' + \left(\frac{u'}{u} + \frac{2}{r}\right) 2k_{\mathrm{B}} T + \frac{\Gamma - \Lambda}{nu} = 0. \tag{10.50}$$

Removing u' via (10.46), we can solve for T':

$$T' = \left[\frac{\Lambda - \Gamma}{nu} - 2k_{\mathrm{B}} T \left(\frac{g}{u} + \frac{2}{r}\right)\right] \times \left[\alpha k_{\mathrm{B}} + \left(\frac{f}{u}\right) 2k_{\mathrm{B}} T\right]^{-1}. \tag{10.51}$$

Given a prescription for Γ and appropriate boundary conditions, (10.51) can be numerically evaluated and thus, using (10.3) and (10.46), the temperature, density, and velocity profiles can be determined.

The radiative cooling term, Λ, includes emission due to magnetic bremsstrahlung, electron–ion and electron–electron thermal bremsstrahlung, and line cooling, but it does not depend on any derivatives (see Melia and Coker 1999 for emissivity expressions). Local UV heating from the massive stars near Sgr A* results in a minimum gas temperature of 10^{4-5} K (Tamblyn *et al* 1996), but the shocked stellar winds at the model's outer radius, R_A, are expected to be hotter than 10^5 K. Thus, the only non-compressive heating term needed in Γ is that

due to magnetic field reconnection. Also, although the expression for Λ depends on the density of ions and electrons, we can here safely assume $n_e = n_i = n$. Following Ipser and Price (1982), for the heating term, we use

$$\Gamma = \frac{nu}{8\pi} \left\{ \left(\frac{B^2}{n}\right)' - \frac{B^2}{n}\left(\frac{u'}{u} - \frac{2}{r}\right) \right\} \tag{10.52}$$

or

$$\Gamma = u\left(\frac{BB'}{4\pi} + \frac{B^2}{2\pi r}\right) \tag{10.53}$$

where B is the magnetic field. If the magnetic field is flux conserved, for which $B(r) \propto r^{-2}$, then no reconnection takes place and $\Gamma = 0$. However, if the density profile is similar to that from adiabatic accretion given in (10.25), then $B(r) \propto r^{-5/4}$, and $\Gamma > 0$. In the case of a dynamo, B increases more rapidly than r^{-2} and $\Gamma < 0$ so that the magnetic field actually cools the gas.

A common assumption is that the magnetic field is in equipartition with some characteristic energy density in the flow. For simplicity we will here assume the magnetic energy density scales with the gravitational potential energy density so that

$$\frac{B^2}{8\pi} = \frac{\delta_B GM\rho}{r} \tag{10.54}$$

where δ_B is a constant scale factor. However, Γ now depends on T' so that (10.51) must be rearranged. With B given by (10.54), (10.52) becomes

$$\Gamma = \frac{\delta_B GM\rho}{r}\left(\frac{u}{r} - fT' - g\right) \tag{10.55}$$

so that

$$T' = \left[\frac{\Lambda}{nu} + \frac{\delta_B GMm_p}{r}\left(\frac{g}{u} - \frac{1}{r}\right) - 2k_B T\left(\frac{g}{u} + \frac{2}{r}\right)\right]$$
$$\times \left[\alpha k_B + \left(\frac{f}{u}\right)\left(2k_B T - \frac{\delta_B GMm_p}{r}\right)\right]^{-1}. \tag{10.56}$$

Given δ_B, M, and boundary values for T, ρ, and u, we can determine the profiles for perfectly convergent supersonic accretion. A representative result for the temperature profile, using values typical for Sgr A*, is shown in figure 10.3 for various values of δ_B. The velocity and density profiles are very close to those for pure Bondi–Hoyle accretion ($r^{-1/2}$ and $r^{-3/2}$ respectively), but the temperature profile deviates significantly from r^{-1} due to the magnetic heating and the change in the equation of state as the electrons become relativistic. Note that with heating and cooling, there are additional length scales present so the solution no longer scales perfectly with the accretion radius.

If the magnetic field is tied to the kinetic or thermal energy density, one gets similar results to figure 10.3. However, based on the fact that magnetic
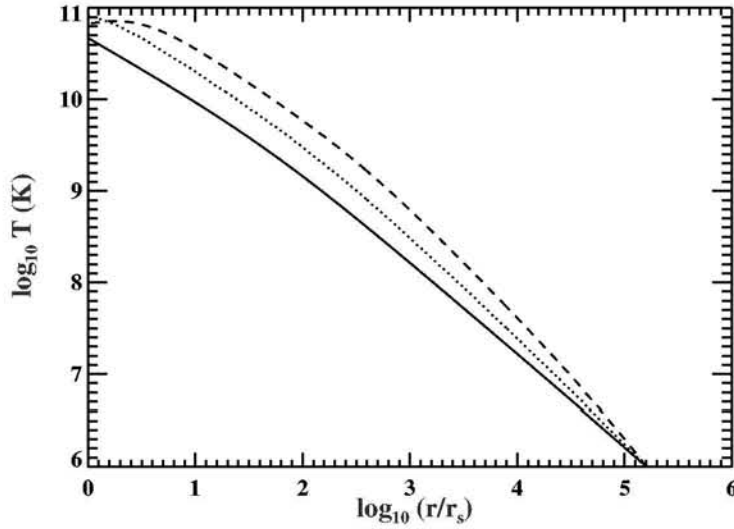
Figure 10.3. Example temperature profiles from solving (10.56). The outer radius of the integration is R_A. Boundary values are $v_\infty = 750\,\text{km s}^{-1}$, $T_\infty = 10^6\,\text{K}$, $\dot{M} = 10^{21}\,\text{g s}^{-1}$. The three curves correspond to $\delta_B = 0$ (full), 0.01 (dotted), and 0.03 (dashed). The black hole mass $M = 2.6 \times 10^6 M_\odot$.

reconnection is not constant so that the field may be subequipartition at large radii and superequipartition at small radii (Melia and Lowalenko 2001), attempts were made to fit the spectrum of Sgr A* using an arbitrary magnetic field profile. The results (Coker and Melia 2000) suggest the flow consists of three regions: a nearly equipartition region at large radii, a region with a flat magnetic field near $\sim 100 r_s$ and a region with a magnetic dynamo near $\sim 4 r_s$. This, combined with hydrodynamical simulations that suggest the accreting gas has sufficient specific angular momentum so that the assumption of spherical accretion breaks down near $\sim 100 r_s$, has led to the conclusion that the spherical accretion model by itself is insufficient to model the accretion flow near Sgr A*.

10.2.3 Radiation from spherical accretion

Compared to active galactic nuclei (AGNs) and even X-ray binaries (XRBs), Sgr A* is emitting very little energy for its mass; i.e. its luminosity is much less than the Eddington luminosity, the critical luminosity above which radiation pressure exceeds gravity. For Sgr A*, the former is $\sim 10^5\ L_\odot$ (Zylka *et al* 1995) while the latter is $\sim 10^{11}\ L_\odot$ (Shapiro and Teukolsky 1983). The appeal of spherical accretion in explaining the spectrum of Sgr A* is that it is inherently inefficient, unlike most other accretion scenarios onto compact objects. For

example, consider the case that the temperature and density scale as r^{-1} and $r^{-3/2}$, respectively, and emission is only due to thermal bremsstrahlung so $\Lambda \propto n^2 T^{1/2}$ (Rybicki and Lightman 1979 equation (5.15b)). If radiative emission is thought of as a minor perturbation of the flow, one can estimate the accretion luminosity as

$$L_{\text{acc}} = \int_{r_{\text{s}}}^{\infty} 4\pi r^2 \Lambda \, dr \tag{10.57}$$

or, in erg s^{-1},

$$L_{\text{acc}} \simeq 10^{31} \left(\frac{T_{\infty}}{10^4 \text{ K}}\right)^{1/2} \left(\frac{n_{\infty}}{1 \text{ cm}^{-3}}\right)^2 \left(\frac{M}{M_{\odot}}\right)^3 \left(\frac{v_{\infty}^2 + c_{\infty}^2}{(1 \text{ km s}^{-1})^2}\right)^{-7/2}. \tag{10.58}$$

For cool, dense gas, line cooling increases the accretion luminosity, while magnetic bremsstrahlung increases the luminosity of an accreting hot magnetized plasma. Assuming just thermal bremsstrahlung, the accretion efficiency is

$$\epsilon = \frac{L_{\text{acc}}}{\dot{M}c^2} \simeq 10^{-4} \left(\frac{n_{\infty}}{1 \text{ cm}^{-3}}\right) \left(\frac{T_{\infty}}{10^4 \text{ K}}\right)^{1/2} \frac{M}{M_{\odot}} \left(\frac{v_{\infty}^2 + c_{\infty}^2}{(1 \text{ km s}^{-1})^2}\right)^{-2} \tag{10.59}$$

where the accretion rate is evaluated at the accretion radius. Thus, for most likely sets of values, $\epsilon \ll 0.1$, the rough value for accretion disks. For the extreme case ($\delta = 0.03$) shown in figure 10.3, there is substantial additional luminosity due to magnetic bremsstrahlung emission, but the accretion efficiency is still only ~4%. Thus, spherical accretion can naturally explain the sub-Eddington luminosity of Sgr A*.

If the mass accretion rate at small radii is near the Bondi–Hoyle estimate and the magnetic field is sufficiently strong, the resulting temperatures and densities at small radii are high enough that the optical depth becomes larger than unity. We can crudely estimate the optical depth, τ, for spherical accretion following Rybicki and Lightman (1979). Since we are particularly interested in τ near the event horizon, we include special and general relativisitic corrections (Shapiro 1973). Thus, $\delta\tau(j) \simeq \delta r \alpha_{\text{abs}}$, with

$$\delta r = \frac{\delta r_0}{\sqrt{(1 - r_{\text{s}}/r)(1 - \beta^2)}} \tag{10.60}$$

where $\delta r_0 = r_{j+1} - r_j$, the observed zone size at infinity, α_{abs} is the absorption coefficient, and β is the bulk velocity of the flow in units of c as measured by a stationary observer:

$$\beta = \frac{u/c}{\sqrt{(u/c)^2 + 1 - r_{\text{s}}/r}}. \tag{10.61}$$

Note that this is not completely correct since we are not including, for example, the capturing of photons by the black hole, which introduces an angular dependence.

For α_{abs}, we use Kirchoff's law,

$$\alpha_{abs} = j_\nu / B_\nu \qquad (10.62)$$

where j_ν is the total emissivity (in erg cm^{-3} s^{-1} Hz^{-1} steradian^{-1}) and B_ν is the blackbody Planck function. For the models discussed here, we use a total emissivity that includes magnetic bremsstrahlung emission (Coker and Melia 1999) and electron–ion and electron–electron thermal bremsstrahlung (Melia and Coker 1999).

In (10.62), ν corresponds to the emitted frequency, not the observed frequency. To find the total optical depth from zone j out to infinity at some observed frequency ν_0, related to the emitted frequency ν by

$$\nu_0 = \nu\sqrt{(1 - r_s/r)(1 - \beta^2)} \qquad (10.63)$$

we use

$$\tau_{\nu_0}^\infty(j) = \sum_{k=j+1}^{k=\infty} (r_j/r_k)^2 \delta\tau(k). \qquad (10.64)$$

The minimum frequency ν_m that a photon needs to have in order to escape from radius r_j is found by determining the frequency at which $\tau_{\nu_0}^\infty(j)$ is unity with the caveat that ν_m not be less than the plasma frequency

$$\nu_p = e\sqrt{\frac{n}{\pi m_e}}. \qquad (10.65)$$

This caveat is required since photons with a frequency less than ν_p are unable to propagate and are thus trapped by the infalling gas. As previously mentioned, it is found that $\nu_m \lesssim 10^{12}$ Hz. Also, at ν_c, the characteristic frequency at which a given radius is primarily emitting, $\tau(r, \nu_c) \gtrsim 1$ only for small radii.

10.2.4 Calculation of the spectrum due to spherical accretion

Once the radial profiles (optical depth, density, velocity, and temperature) are determined, it is possible to calculate the emission spectrum for a given magnetic field profile. The predicted observable luminosity L_{ν_0} at infinity (Shapiro 1973, Melia 1992) with relativistic corrections is

$$L_{\nu_0} = 16\pi^2 \sum_{j=1}^{j=J} e^{-\tau_{\nu_0}^\infty(j)} r_j^2 (1 - \beta^2)^{3/2} (1 - r_s/r_j) I_\nu \qquad (10.66)$$

where

$$I_\nu = B_\nu(1 - e^{-\delta\tau(j)}). \qquad (10.67)$$
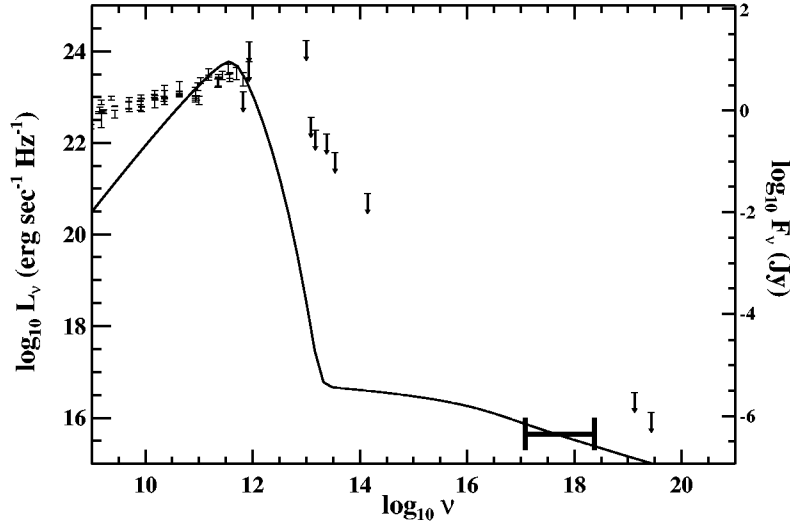
Figure 10.4. The full curve is an example spectrum profile from solving (10.56) and (10.66) with parameters as in figure 10.3 but with $\delta_B = 0.001$. Also shown are the observed values and upper limits with the recent Chandra results (Baganoff *et al* 2001a, b) highlighted in bold (see chapter 11 for a discussion of the observed spectrum). At a distance of 8.5 kpc, for Galactic Center sources $F_\nu \simeq 10^{23} L_\nu$.

The sum over j is truncated at J, for which $r_J = R_A$. It is assumed that $\tau_{\nu_0}^\infty(J) = 0$. This ignores the possible absorption by Sgr A West of the low-frequency ($\nu_0 < 10^9$ Hz) radiation (Beckert *et al* 1996). Sgr A West is an H II region surrounding Sgr A*.

An example of a spectrum arising from these equations is shown in figure 10.4. Although only a representative solution, the primary features of all spherical accretion models are apparent. First, there is little emission in the infrared, unlike disk accretion scenarios where the gas circularizes and thermalizes before accreting. Second, there is significant X-ray emission due to thermal bremsstrahlung. However, in the case of Sgr A*, the observed quiescent X-ray spectrum is too soft to be due to thermal bremsstrahlung alone; the sub-mm emission, coming from very close to the black hole, is likely to be upscattered via inverse Compton and produce additional X-ray emission. Third, the spectral index in the radio is ∼1, rather steeper than the observed value of ∼0.3 (Falcke *et al* 1998). This last characteristic deserves some attention.

If the gas temperature and density profiles are at all similar to the Bondi–Hoyle results given earlier, and, if the radio emission is due to magnetic bremsstrahlung from a thermal distribution of particles, it is impossible to match both the observed 1 GHz radio flux and the soft X-ray upper limits (Liu and

Melia 2001). Since the emission must be blackbody limited, one can put a lower limit on the temperature at a given radius. Similarly, the gas temperature must be less than the virial temperature, putting an upper limit on the temperature. For frequencies less than $h\nu/k_{\mathrm{B}}T$, the thermal bremsstrahlung emissivity, j_ν^b, scales as $n^2 T^{-1/2} \propto r^{-5/2}$. Thus, the volume-integrated emissivity will increase with the size of the emission region as $r^{1/2}$ so that the minimum luminosity corresponds to the minimum size and maximum temperature consistent with the blackbody and virial limits. Limits which result in the proper observed 1 GHz flux yield a minimum radius and maximum temperature of approximately $2000\,r_s$ and 10^9 K, respectively. Since the soft X-ray emission is not likely to be self-absorbed while the GHz emission may be, the ratio of the volume integrated emissivities must be less than the ratio of the observed fluxes. For an equiparition-type magnetic field profile, the result is that the soft X-ray luminosity cannot be less than $\sim 10^{-5}$ times the 1 GHz luminosity. Thus, since the observed ratio is $\sim 10^{-7}$ (Falcke *et al* 1998, Baganoff *et al* 2001a, b), the X-ray emission cannot be due to thermal magnetic bremsstrahlung. Profiles which deviate substantially from the Bondi–Hoyle results can avoid this problem. If one assumes an inflow velocity that is larger than free-fall, for example, one can reproduce the radio and X-ray emission, even when including self-Compton (Coker and Markoff 2001). However, the most likely cause of deviation from Bondi–Hoyle is a non-spherical accretion flow so that one has effectively a radially dependent mass accretion rate. A non-thermal particle distribution due to shocks in the flow could also alter the spectrum enough to invalidate these arguments.

10.3   Non-spherical accretion models

Both observationally and theoretically the bulk of the radio emission from Sgr A* appears to come from close to the event horizon. At small radii the assumption of spherical symmetry is probably invalid. Therefore, we would like to find self-consistent solutions to the equations of motion which include differentially rotating gas flows. We discuss two such models here: the Keplerian flow dynamo model and the sub-Eddington two-temperature accretion model. Other commonly used models, such as the well-known thin-disk model (Shakura and Sunyaev 1973) probably do not apply to Sgr A* since they predict substantial infrared emission which is not seen in the Galactic Center. In fact, combined with the presence of the stellar winds, it seems unlikely that any true large-scale disk exists around Sgr A* (Coker *et al* 1999). However, the sub-Eddington two-temperature accretion model discussed in section 10.3.2 results in a disklike accretion flow; it has been argued that this flow incorporates the stellar winds without either emitting significant infrared or being destroyed (Narayan *et al* 1998).

This apparent absence of a disk and any associated jet sets Sgr A* apart from many other black hole systems such as luminous AGNs and XRBs. The stars that feed Sgr A* are fairly uniformly distributed around the black hole so

that it is probably accreting relatively little angular momentum. The direction of the accreting angular momentum vector is likely to be time variable as well. Thus, the complete picture probably requires a combination of large-scale spherical accretion with a small-scale (and as yet unobserved) disk and/or jet.

A rough example of what the large-scale flow might look like near Sgr A* is shown in figure 10.5. The 10 wind sources in this hydrodynamical model (see Coker and Melia 1997 for details) produce large-scale shock fronts and cavities with time-dependent characteristics. In reality there are at least two dozen (Genzel *et al* 1996) stellar sources and some of them may be $\lesssim 1$ R_A from Sgr A*. Also, the stars move relative to one another while in this simulation they are stationary. Thus, the flow is likely to be even more highly non-spherical than is shown in the figure. In fact, if the sources are rotating as a cluster (Genzel *et al* 2000), the flow may have sufficient angular momentum to circularize at a radius as large as $10^4 r_s$.

10.3.1 Keplerian flow with magnetic dynamo

This model assumes the sub-mm and X-ray emission of Sgr A* arises from the circularization of the infalling gas at very small radii (Melia *et al* 2001). We construct a standard accretion disk with the assumption that turbulence produces a magnetic field that is predominantly azimuthal. We start with the solution to the Euler equation for inward viscous transport (Shakura and Sunyaev 1973) in a Keplerian disk,

$$2\rho H v_r \frac{d(\Omega r^2)}{dr} = \frac{1}{r} \frac{d(W_{r\phi} r^2)}{dr} \tag{10.68}$$

where v_r is the (positive inward) radial velocity of the gas, H is the height of the disk,

$$\Omega = \left(\frac{GM}{r^3}\right)^{1/2} \tag{10.69}$$

is the Keplerian rotational frequency, and $W_{r\phi}$ is the vertically integrated stress. We assume that $v_r \ll r\Omega$ for all r. The disk height is found by balancing gravity with the vertical pressure gradient

$$H = c_s/\Omega = \sqrt{\frac{2k_B T r^3}{\mu m_p GM}} \tag{10.70}$$

where c_s is the isothermal sound speed. The stress is related to the kinematic shear viscosity v by

$$W_{r\phi} = 3v\rho H \Omega. \tag{10.71}$$

We assume the stress is dominated by the Maxwell stress so that if we assume the magnetic energy density scales with the thermal energy density, we have

$$W_{r\phi} = \frac{C\rho H k_B T}{\mu m_p} \tag{10.72}$$
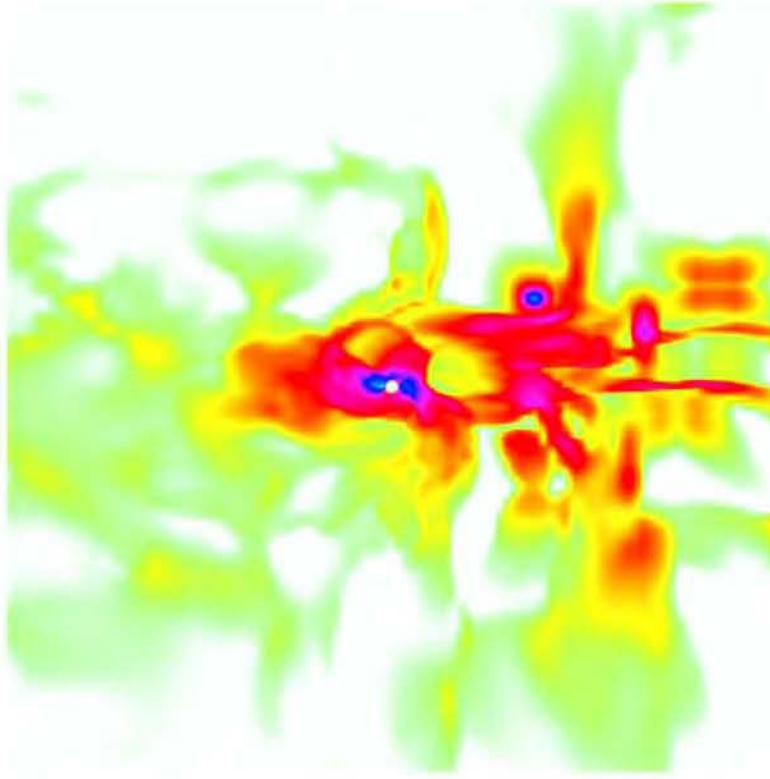
Figure 10.5. A plot of logarithmic density through the $\hat{x}$–$\hat{z}$ plane for a sample hydrodynamical simulation applicable to Sgr A*. The image is 16 $R_A \simeq 0.3$ pc $\simeq 10^6 \, r_s$ on a side. The darkest color corresponds to a number density $\gtrsim 10^4$ cm^{-3} and white corresponds to a number density $\lesssim 10^2$ cm^{-3}. Sgr A* is modeled as a perfectly absorbing sphere with a radius of 0.1 R_A. The 10 wind sources, one of which is visible to the upper right of Sgr A*, have been blowing for ~2500 years. See also color section.

where C is a constant scale factor. Numerical MHD simulations suggest $C \sim 0.01$ (Brandenburg *et al* 1995). Note that with these definitions, $C = 3\alpha$, where α is the standard Shakura–Sunyaev viscosity parameter (see section 10.3.2). We also have the disk version of the mass continuity equation

$$\dot{M} = 4\pi r \rho H v_{\rm r}. \tag{10.73}$$

Together, this permits us to write the integral (10.68) as

$$v_{\rm r} = \frac{r T(r)}{T_0 r_0 / v_{r0} - (GM)^{1/2}(r_0^{1/2} - r^{1/2})(\mu m_{\rm p} / C k_{\rm B})} \tag{10.74}$$

where the quantities with subscript 0 are evaluated at the outer edge of the disk (i.e. at radius r_0). For (10.74) to represent a physical accretion solution, the denominator of (10.74) must not reach zero. Letting $v_r \rightarrow \infty$ at the inner edge of the disk ($\equiv 1 \ r_s$), we get T_0 in terms of r_0 and v_{r0}:

$$T_0 = \frac{v_{r0}}{r_0} \frac{m_p \mu (GM)^{1/2}(r_0^{1/2} - r_s^{1/2})}{C k_B}. \tag{10.75}$$

Turbulence with perturbing wavenumbers $k \simeq \Omega/v_{Az}$, where v_{Az} is the Alfven velocity due to the vertical magnetic field, is unstable and generates a positive feedback loop between the kinetic energy density of the turbulence and the energy density of the turbulent magnetic field. This 'dynamo' turns off when the two energy densities are in rough equipartition. The net result is that the equilibrium magnetic energy density is a fraction $\beta_B \sim 0.03$ of the total thermal energy density. Many numerical simulations have verified this expectation (Hawley *et al* 1995). Also, the most unstable modes are not damped by Ohmic diffusion because the diffusion length for gas $\sim 10 \ r_s$ from Sgr A* is extremely small (Melia *et al* 2001). Given the disparate stellar sources of the infalling gas, a dominant large-scale ordered magnetic field is unlikely. In such a case, the turbulent magnetic field will likely dominate the final magnetic energy density (but see the discussion in section 10.2.2).

It remains to solve the total (magnetic, kinetic, and thermal) energy equation in order to derive the temperature profile. This requires solving the equation for thermal energy conservation as well as the magnetic evolution equation in the presence of Ohmic diffusion and viscous dissipation (Balbus and Hawley 1991):

$$\frac{\partial \epsilon}{\partial t} + \nabla \cdot (v\epsilon) - \Gamma + \Lambda + P\nabla \cdot v = 0 \tag{10.76}$$

$$\frac{\partial B}{\partial t} = \nabla \times (v \times B) - \eta \nabla \times (\nabla \times B) \tag{10.77}$$

where η is the resistivity of the fully ionized gas (Spitzer 1962). We assume azimuthal symmetry and incompressibility and neglect radiation pressure and buoyancy in order to find the steady-state solutions to (10.76) and (10.77). Since we are not using the relativistic Euler equation, the results will not be strictly valid near r_s; more complicated corrections such as those due to light bending are also ignored.

The most difficult aspect of determining the accretion profiles is solving for the steady-state dissipative heating term, Γ, in terms of the current density and stress. This requires ignoring the divergence of the viscous and Ohmic fluxes and taking only the high-order terms. More details are given in Melia *et al* (2001) but the resulting differential equation for the temperature is

$$E_1 T' = E_2 + \frac{\Lambda m_p}{\rho v_r k_B} \tag{10.78}$$

where

$$E_1 = \alpha + 2 + \frac{2\beta_B}{\mu} - \frac{3Cr\Omega}{2\mu v_r} - \frac{m_p v_r^2}{2k_B T} + \frac{3}{2} E_3$$

$$E_2 = \frac{2T Cr\Omega}{r \mu v_r} + \left(\frac{5}{2r} v_r^2 - \frac{GM}{2r^2} \right) \frac{m_p}{k_B} - E_3 T \left(\frac{7}{2r} - \frac{m_p \mu \sqrt{GM} v_r}{2Ck_B T} r^{-3/2} \right)$$

$$E_3 = \frac{m_p v_r^2}{k_B T} + \frac{Cr\Omega}{\mu v_r}. \tag{10.79}$$

Together with (10.73) and (10.74), equation (10.78) determines the accretion profile, provided some boundary conditions are given. An example of some profiles that may be applicable to Sgr A* are shown in figure 10.6. Note that while spherical accretion is always stable (Moncrief 1980), it has not been determined whether the disklike profiles given here produce a thermally or dynamically stable disk.

10.3.1.1 Calculation of the spectrum

We calculate the spectrum due to this Keplerian flow model in a fashion that parallels section 10.2.4. At a frequency v_0, the predicted observed flux density produced by the Keplerian flow is given by

$$F_{v_0} = \frac{1}{D^2} \int I_{v_0} \sqrt{1 - r_s/r} \, \mathrm{d}A \tag{10.80}$$

where $D = 8.5$ kpc is the distance to the Galactic Center and the observed frequency at infinity is now

$$v_0 = v \frac{\sqrt{(1 - r_s/r)(1 - \beta^2)}}{1 - \beta \cos \theta}. \tag{10.81}$$

The $\sqrt{1 - r_s/r}$ term in (10.80) is due to the gravitational redshift. The frequency measured in the comoving frame is v,

$$\beta \equiv \beta_\phi = \frac{r\Omega/c}{\sqrt{(r\Omega/c)^2 + 1 - r_s/r}} \tag{10.82}$$

is the normalized azimuthal velocity seen by a stationary observer, and θ is the angle between $\boldsymbol{\beta}$ and the line of sight. Thus,

$$\cos \theta \equiv \sin i \cos \phi \tag{10.83}$$

where i is the inclination angle with respect to the line of sight of the rotation axis perpendicular to the Keplerian flow, and ϕ is the position angle of the emitting element. Figure 10.7 shows a cartoon of the accretion region. The area element is

$$\mathrm{d}A = \frac{\cos i \, r \, \mathrm{d}r \, \mathrm{d}\phi}{\sqrt{1 - r_s/r}} \tag{10.84}$$
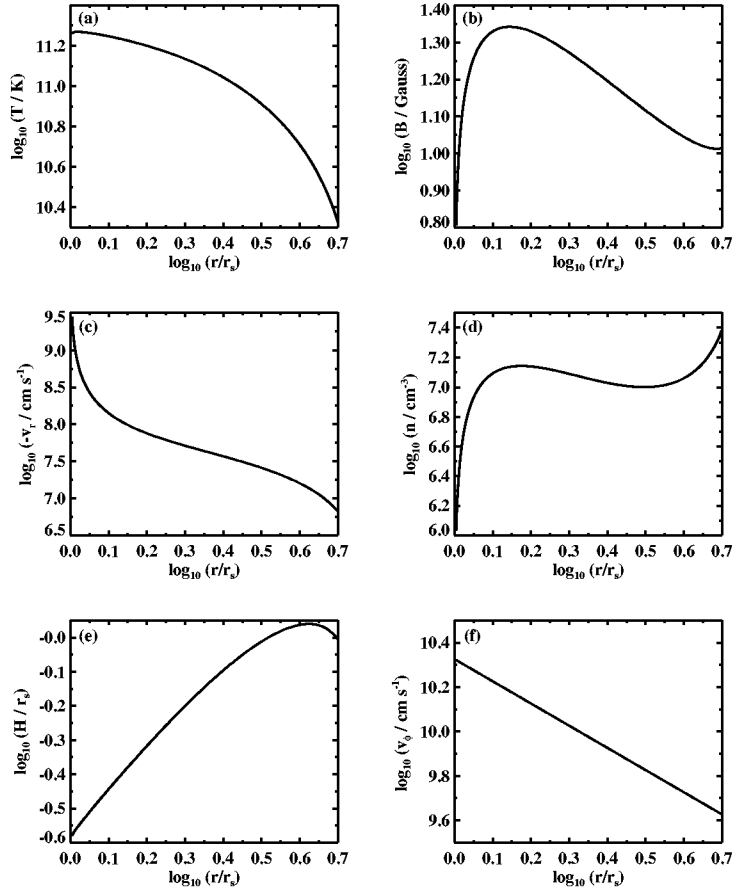
Figure 10.6. Example profiles for a Keplerian accretion with magnetic dynamo model for Sgr A*. Here $\dot{M} = 10^{16}$ g s^{-1}, $v_{r0} = 5 \times 10^{-4}$ v_{kep}, $C = 0.01$, $\beta_B = 0.03$, and $r_0 = 5\,r_s$.

and the specific intensity

$$I_{\nu_0} = B_{\nu_0}(1 - e^{-\tau})$$
(10.85)

where

$$B_{\nu_0} = \left(\frac{\sqrt{(1 - \beta^2)(1 - r_s/r)}}{1 - \beta \cos\theta} \right)^3 B_\nu$$
(10.86)

and B_ν is again the blackbody Planck function. The optical depth from the emitting element to the observer is approximately

$$\tau = \int \alpha_\nu \, ds = \frac{j_\nu}{B_\nu} \frac{2H}{\cos i} \frac{1 - \beta \cos\theta}{\sqrt{(1 - \beta^2)(1 - r_s/r)}}$$
(10.87)
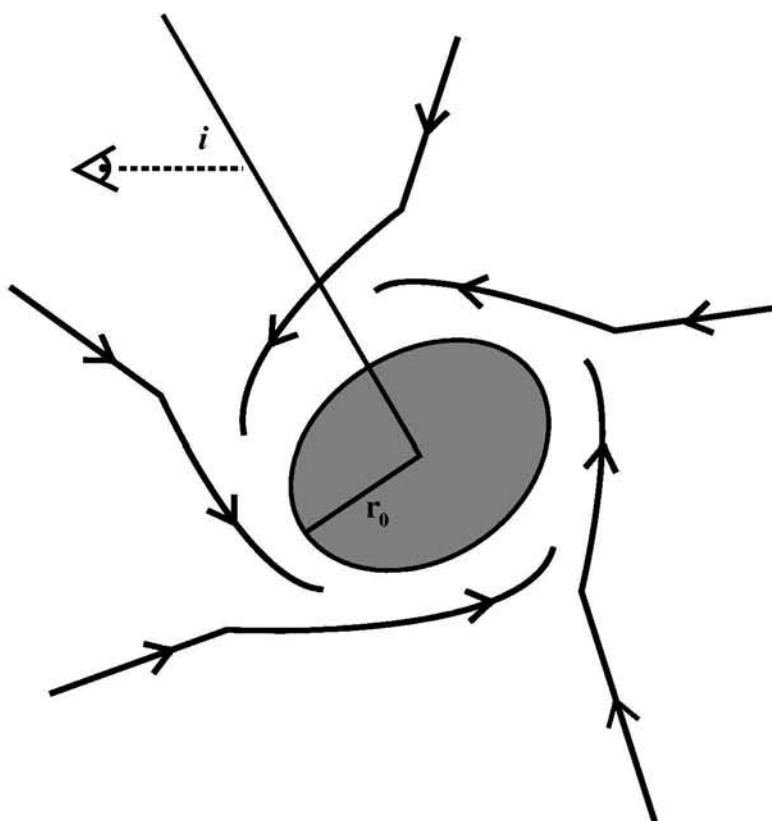
Figure 10.7. Schematic diagram showing streamlines of infalling gas for the Keplerian accretion model.

where j_ν is the emissivity in the comoving frame. Note that the transformations in these equations implicitly assume that $\beta_\phi \gg \beta_r$ for all r. Given a prescription for the emissivity, this determines the spectrum.

There are some differences between the derivation here and in section 10.2.4. The geometry of a disk instead of spherical inflow results in a different treatment of the optical depth. In a disk there is only local absorption as in (10.85) while in spherical flow a photon can be absorbed at large radii as can be seen in (10.66). Also, here we can no longer assume that $\cos \phi = 0$. Instead, it is necessary to explicitly integrate over ϕ.

An example of a spectrum, assuming an emissivity due to relativistic magnetic bremsstrahlung from thermal particles (Pacholczyk 1970), is given in figure 10.8. The input parameters are the same as for the profiles shown in figure 10.6. With some tuning, this model can reproduce the sub-mm 'bump' in the radio emission but has too steep a spectral index, resulting in very little
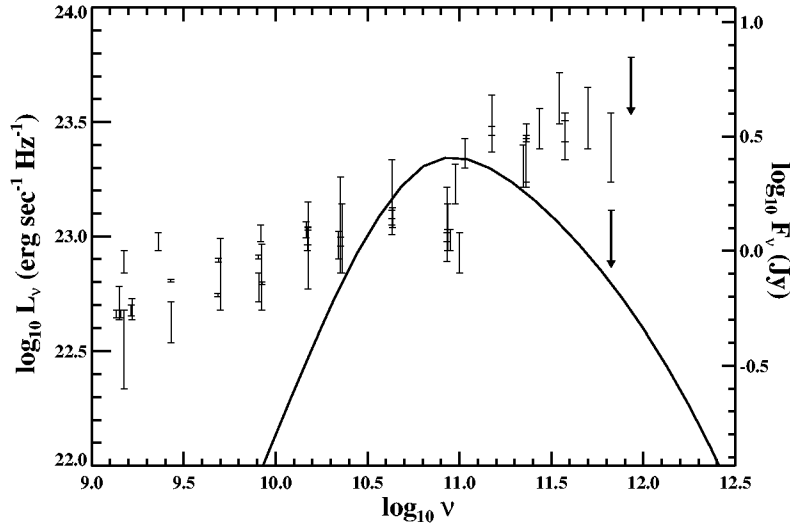
Figure 10.8. Example spectrum for a Keplerian accretion model for Sgr A* using the profiles shown in figure 10.6. Here we use a representative disk inclination of 60°.

low-frequency emission. As previously stated, this is expected for a thermal distribution of particles and implies that the low-frequency emission is distinct from the sub-mm. Perhaps the low-frequency radio emission originates in a small jet (see chapter 11) or further out in the semi-spherical accretion flow. The X-ray emission due to thermal bremsstrahlung from the model shown in figure 10.8 is considerably less than the detected X-ray flux. However, by including inverse Compton, which upscatters the sub-mm-wave photons, one can reproduce the observed flux and spectral index of the X-ray emission. The primary conclusion of this model is that in order to produce consistent radio emission without violating the IR or X-ray limits, one requires a mass accretion rate at $\sim 10 \, r_s$ that is around three orders of magnitude less than that predicted from the simple Bondi–Hoyle estimate. Given the good constraints on the large-scale winds, this implies a strongly radially dependent mass accretion rate. Also, the lower mass accretion rate results in a higher final radiative efficiency of $\simeq 7\%$, about that expected for accretion disk scenarios.

10.3.2 Sub-Eddington two-temperature accretion (ADAFs)

As in the previous section, we consider an axisymmetric steady-state disk accretion flow. The mass continuity equation and disk scale height are still given by (10.73) and (10.70). However, the azimuthal velocity is no longer assumed to

be given by the Keplerian velocity so that

$$v_\phi \equiv r\omega \neq \sqrt{\frac{GM}{r}}. \tag{10.88}$$

Thus, the non-relativistic equation for the radial velocity now has an extra centrifugal force term compared to (10.37):

$$v_r v_r' + \frac{1}{\rho} P' - \omega^2 r + \frac{GM}{r^2} = 0. \tag{10.89}$$

For simplicity, we will deal only with the non-relativistic equations in this section. Again, v_r is defined positive for inflow. In this model, we ignore the dynamical effects of the magnetic field. The shear viscosity, ν, is still given by (10.71) and (10.72) but is now written as

$$\nu \equiv \alpha c_s H = \alpha c_s^2 / \Omega \tag{10.90}$$

where α is the r-independent α-viscosity parameter (Shakura and Sunyaev 1973) and now c_s is the isothermal sound speed ($c_s^2 \equiv P/\rho = 2k_B T/m_p$). Note that in section 10.3.1, $\alpha \simeq 0.003$. This viscosity causes a change in ω with r which follows from the vertically integrated form of the azimuthal component of the Euler equation:

$$v_r \rho r H \frac{d(\omega r^2)}{dr} + \frac{d}{dr}\left(\nu \rho r^3 H \frac{d\omega}{dr}\right) = 0. \tag{10.91}$$

Compare (10.91) with the less general (10.68); both equations assume that the flow is viscous and therefore reduce to unphysical results ($\omega \propto r^{-2}$) in the limit of $\nu \to 0$. Using the first law of thermodynamics, written as

$$m_p T \, ds = \frac{k_B \, dT}{\gamma - 1} - k_B T \frac{dn}{n} \tag{10.92}$$

where s is the specific entropy and assuming a constant γ and $\mu = 0.5$, we can rewrite (10.50) as

$$m_p T s' = \frac{2k_B T'}{\gamma - 1} - \frac{2k_B T n'}{n} = \frac{\Lambda - \Gamma}{n v_r} \tag{10.93}$$

where again a prime ($'$) denotes d/dr. Here, we ignore magnetic heating and instead assume that the heating is purely due to an anomalous shear viscosity so that

$$\Gamma = \rho \nu r^2 \left(\frac{d\omega}{dr}\right)^2. \tag{10.94}$$

It is the ratio $\eta \equiv \Gamma/\Lambda$ which determines the type of accretion flow: $\eta \simeq 1$ corresponds to the standard thin-disk solution, $\eta \gg 1$ corresponds to the flow

discussed here, and $\eta \ll 1$ corresponds to Bondi-type solutions. All three types may be present near Sgr A*. In this section, we are interested in solutions in which the viscous heating rate Γ is much greater than the radiative cooling rate Λ; for the Keplerian model discussed in section 10.3.1, $\Gamma/\Lambda \ll 1$ except very close to the outer edge of the flow. The $\eta \gg 1$ solutions, it turns out, require sub-Eddington mass accretion rates and a two-temperature plasma with hot ions and cool electrons. These models are known as Advection Dominated Accretion Flows (ADAFs; see Narayan *et al* 1998) because the ions of the gas are heated by viscous dissipation and accretion occurs before the cooler electrons can radiate. Bondi–Hoyle-type solutions also advect most of their energy without radiating (thus the low radiative efficiency described earlier) so the term ADAF is slightly misleading.

If we further assume (Ichimaru 1977, Narayan and Yi 1994, Narayan *et al* 1998) that $\Lambda \to 0$ and α is constant with r, we find a self-similar solution exists. That is, n, T, v, and ω can be given by simple power laws: $\rho \propto r^{-3/2}$, $v \propto r^{-1/2}$, $T \propto r^{-1}$, and $\omega \propto r^{-3/2}$. In particular, if $A \equiv \omega/\Omega$ and the Keplerian velocity is

$$v_{\mathrm{K}} \equiv r\Omega = \sqrt{\frac{GM}{r}} \tag{10.95}$$

then (10.89), (10.91), and (10.93) become

$$\frac{v_{\mathrm{r}}^2}{2} + (A^2 - 1)v_{\mathrm{K}}^2 + \frac{5}{2}c_{\mathrm{s}}^2 = 0 \tag{10.96}$$

$$v_{\mathrm{r}} = \frac{3\alpha c_{\mathrm{s}}^2}{2v_{\mathrm{K}}} \tag{10.97}$$

and

$$c_{\mathrm{s}}^2 = \frac{(\gamma - 1)A^2 v_{\mathrm{K}}^2}{5/3 - \gamma}. \tag{10.98}$$

Solving these three equations and (10.73) and taking the limit $\alpha \ll 1$ gives:

$$v_{\mathrm{r}} = v_{\mathrm{K}}\frac{\gamma - 1}{\gamma - 5/9}\alpha \tag{10.99}$$

$$c_{\mathrm{s}}^2 = \frac{2(\gamma - 1)v_{\mathrm{K}}^2}{3(\gamma - 5/9)} \tag{10.100}$$

$$\omega = \Omega\sqrt{\frac{2(5/3 - \gamma)}{3(\gamma - 5/9)}} \tag{10.101}$$

and

$$\rho = \frac{\dot{M}}{\alpha 4\pi r^3 \Omega}\sqrt{\frac{3(\gamma - 5/9)}{2(\gamma - 1)}}. \tag{10.102}$$

These results are similar to the profiles derived for spherical Bondi-type accretion.

In most ADAF models, it is generally assumed that a perfectly tangled magnetic field exists with a magnetic pressure that is comparable to the thermal pressure. A purely tangled field results in an added specific internal energy density term of $B^2/8\pi\rho$ and a radial pressure term of $B^2/24\pi$; this pressure is assumed to be a constant fraction $(1 - \beta_B)$ of the total (thermal plus magnetic) pressure (note the difference in definition of this β_B from that in section 10.3.1). Then we have

$$P_{\text{tot}} = \frac{B^2}{24\pi} + P_{\text{th}} = \frac{B^2}{24\pi} + \beta_B \rho c_s^2 = \frac{B^2}{24\pi} + n k_B \left(\frac{T_i}{\mu_i} + \frac{T_e}{\mu_e} \right) \quad (10.103)$$

where P_{th} is the thermal component of the pressure and T_i and T_e are the ion and electron temperatures (assumed equal at this point). For simplicity we assume a hydrogen mass fraction of one so that the mean molecular weights of both the ions or protons, μ_i, and electrons, μ_e are unity. The magnetic field changes the effective value of γ so that if it is assumed that the hydrodynamical adiabatic index, $\gamma_{\text{ad}} = 5/3$, then

$$\gamma = \frac{8 - 3\beta_B}{6 - 3\beta_B}. \quad (10.104)$$

This treatment of the field is not fully accurate since even a tangled field will also result in some resistance to the shear viscosity. In addition, the Balbus–Hawley magneto-rotational instability (Balbus and Hawley 1991) will probably eventually result in a magnetic field with a dominant azimuthal component. In the following, $\beta_B = 0.5$.

A problem with self-consistency arises at this point: do the profiles given by (10.99)–(10.102) actually result in $\Gamma \gg \Lambda$? Substituting the self-similar results into (10.94) gives

$$\Gamma = \sqrt{\frac{3(\gamma - 1)}{2(\gamma - 5/9)} \frac{3\dot{M}GM}{8\pi r^4}}. \quad (10.105)$$

Similarly, assuming cooling is due to thermal bremsstrahlung (Rybicki and Lightman 1979) gives (in cgs units):

$$\Lambda = \frac{6.7 \times 10^{16}}{32\pi^2\alpha^2} \sqrt{\frac{3(\gamma - 5/9)}{\gamma - 1}} \frac{\dot{M}^2}{\sqrt{GM}} r^{-3/2}. \quad (10.106)$$

Therefore we get:

$$\eta = 7 \times 10^{-27} \frac{\gamma - 5/9}{\gamma - 1} \frac{\alpha^2 M^{3/2}}{r^{1/2}\dot{M}}. \quad (10.107)$$

We parametrize the mass accretion rate with $f \equiv \dot{M}/\dot{M}_E$, so that $\dot{M} = f \times 10^{24}M_6$ g s^{-1}, where $\dot{M}_E$ is the accretion rate that, with 10% efficiency, produces the Eddington-limited luminosity (see section 10.2.3), and M_6 is the mass of the black hole in units of $10^6 M_\odot$. Then in order to have an ADAF solution, we must have

$$f \ll 10^{-7}\alpha^2 r_0^{-1/2} \quad (10.108)$$

where here r_0 is the outer boundary of the flow in units of r_s. If one includes magnetic bremsstrahlung and Compton cooling, f must be even smaller than the limit given in (10.108). Therefore, it is unlikely that large-scale single-temperature ADAFs exist except in rare circumstances where the accretion rate is particularly low.

The usual way around this difficulty is to assume that the cooling is *not* given by (10.106) because the temperature of the radiating electrons, T_e, is not the same as that of the bulk of the gas. It is argued that viscous heating applies primarily to the massive ions and the Coulomb coupling between the hot ions and cool electrons is sufficiently weak to result in a so-called 'two-temperature plasma'. Ideally, such a problem should be solved using a full two-fluid approach, but for simplicity we shall just assume that $\Lambda \rightarrow 0$ for the ions and solve explicitly for the ion and electron temperatures. Thus, instead of (10.93), T_i is given by (Esin *et al* 1997)

$$\left(\frac{3(1 - \beta_B)}{\beta_B} + \alpha(\theta_p) + T_i \frac{d\alpha(\theta_p)}{dT_i} \right) T_i' = \frac{T_i n'}{\beta_B n} - \frac{\Gamma \mu_i}{k_B n v_r} \qquad (10.109)$$

and

$$\left(\frac{3(1 - \beta_B)}{\beta_B} + \alpha(\theta_e) + T_e \frac{d\alpha(\theta_e)}{dT_e} \right) T_e' = \frac{T_e n'}{\beta_B n} + \frac{(\Lambda - \Gamma m_e/m_p - \Gamma^{ie})\mu_e}{k_B n v_r} \qquad (10.110)$$

where

$$\alpha(x) = x \left(\frac{3K_3(x) + K_1(x)}{4K_2(x)} - 1 \right). \qquad (10.111)$$

After some manipulation, it can be shown that

$$\frac{d\alpha(\theta_e)}{dT_e} = \frac{(1 + \alpha)(3 - \alpha)}{T_e} + \frac{(3 - 2\alpha)\theta_e}{T_e}. \qquad (10.112)$$

The heating of the electrons due to Coulomb coupling with the ions, Γ^{ie}, is given by (Stepney and Guilbert 1983)

$$\Gamma^{ie} = 30n^2 \sigma_T c \frac{m_e}{m_p} \frac{k_B T_i - k_B T_e}{K_2(1/\theta_e) K_2(1/\theta_p)} \left(\frac{2(\theta_e + \theta_p)^2 + 1}{\theta_e + \theta_p} K_1(\zeta) + 2K_0(\zeta) \right). \qquad (10.113)$$

The K_i are the modified Bessel functions, σ_T is the Thomson cross-section, $\theta_e \equiv k_B T_e/m_e c^2$, $\theta_p \equiv k_B T_i/m_p c^2$, and $\zeta \equiv (\theta_e + \theta_p)/\theta_e \theta_p$. It is assumed in (10.113) that the Coulomb logarithm is $\simeq 20$. Since it is assumed that $T_i > T_e$ everywhere, there is no direct feedback from the electrons to the ions; since $\rho c_s^2 \equiv P_{tot}$, there is indirect feedback. The profiles for the gas radial velocity, rotational frequency, and density are still given by (10.99), (10.101), and (10.102), respectively but with $T = T_i$. Note that in earlier work (Narayan and Yi 1995a)
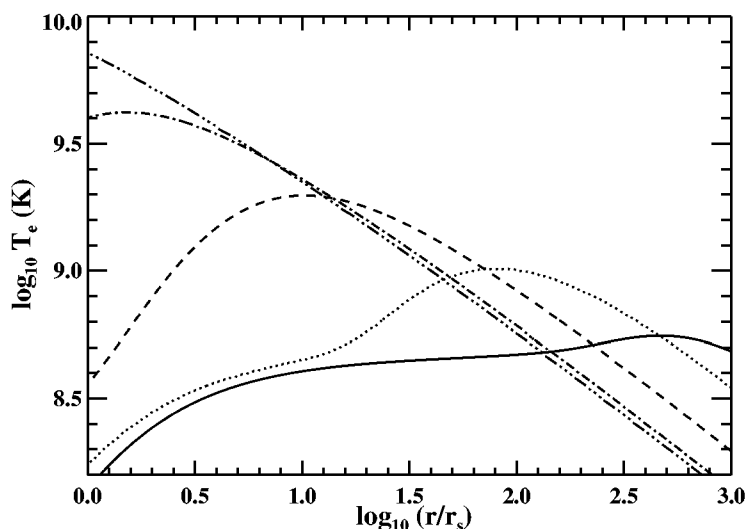
Figure 10.9. Examples of electron temperature profiles for a two-temperature ADAF. Here, $\beta_B = 0.5$, $\Lambda_i = 0$, $\alpha = 0.1$, $\gamma_{ad} = 5/3$, and $r_o = 10^5 r_s$. The curves correspond to different values of $\log_{10}(f)$: -4 (full), -5 (dotted), -6 (dashed), -7 (dot–dashed), and -8 (chain dot–dashed).

the electrons were assumed to be in isentropic balance so that the left-hand side of (10.110) was set to zero.

The existence of a two-temperature ADAF solution now depends on the Coulomb coupling timescale compared to the accretion timescale. A detailed calculation shows that as long as $f \lesssim 0.3\alpha^2$, a two-temperature ADAF solution exists. Representive profiles for T_e are shown in figure 10.9 with $\alpha = 0.1$. For this plot, the assumption that $\alpha \ll 1$ has been relaxed. Also, radiation pressure and the optical depth are assumed to be negligible. The cooling function for the electrons includes magnetic bremsstrahlung, thermal bremsstrahlung (both electron–electron and electron–ion), and line cooling but not inverse Compton scattering. For large accretion rates, the last could result in significant cooling within $\lesssim 10$ r_s (Mahadevan 1997). For comparison, the ion temperature goes as r^{-1} and $T_i \simeq 10^{10}$ K at $r = 100$ r_s. The flow is subsonic everywhere with a Mach number of ~ 0.07; for comparison, the Keplerian model in section 10.3.1 has a Mach number of ~ 0.5. Thus, both models have implicitly assumed that the flow goes supersonic infinitesmally close to the event horizon. More advanced numerical ADAF models have addressed this issue with more appropriate inner boundary conditions (Narayan *et al* 1997, Gammie and Popham 1998).

Since the gas flow in an ADAF is fairly similar to the Keplerian flow described in section 10.3.1, the prescription for determining the spectrum detailed
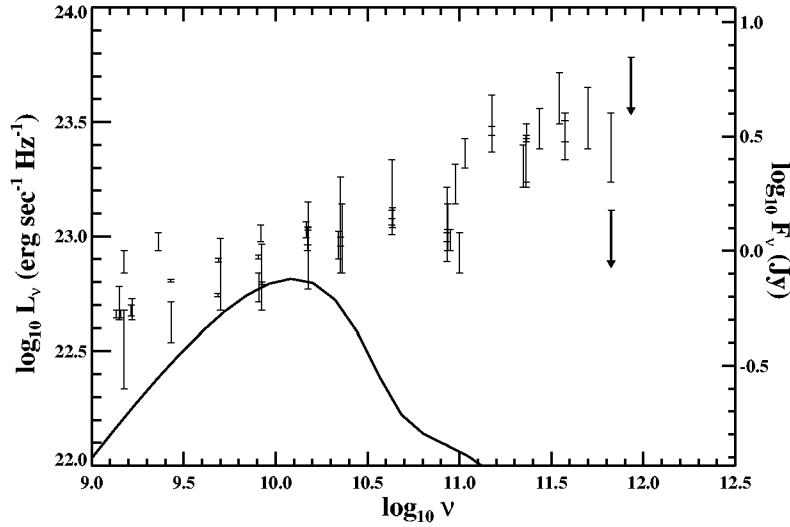
Figure 10.10. Example spectrum for a two-temperature ADAF accretion model for Sgr A* using the profiles shown in figure 10.9. Here we use representative values of $\alpha = 0.3$ and $f = 10^{-4}$. The flow is inclined 60° to the line of sight.

in section 10.3.1.1 can be used if one assumes the ions play no role in the emission. The resulting spectrum for $\alpha = 0.3$ and $f = 10^{-4}$ is given in figure 10.10. The accretion rate for this model is $\dot{M} = 4 \times 10^{20}$ g s^{-1}, much larger than the accretion rate used in the Keplerian model which produces the spectrum shown in figure 10.8. The lower electron temperature results in less magnetic bremsstrahlung (even though the magnetic field for an ADAF is stronger since it is tied to the high-temperature ions). Also, for this representative case, the X-ray spectrum due to thermal bremsstrahlung from large radii is somewhat higher than the observations and in any case does not match the spectral index. We have ignored the fact that the magnetic field in an ADAF is assumed to be purely tangled while the field in the Keplerian flow is assumed to be purely azimuthal; the same magnetic bremsstrahlung emissivity has been used.

10.3.2.1 Comments on ADAFs

Except for the requirement that there be a large amount of angular momentum in the accreting gas, the properties of the flow at infinity have little impact on the structure and spectrum of ADAFs, making for very straightforward modeling. With fixed ion–electron coupling and viscosity, the primary free parameter is the mass accretion rate. The flow is purely determined by shear viscosity. In fact, it is explicitly assumed that viscosity is important, or else $v_r \to 0$ and no accretion

occurs. Also, it has been pointed out (Narayan and Yi 1995a) that the Bernoulli parameter of ADAFs is positive, meaning that the gas has positive energy at infinity. Thus, it is possible that ADAFs occur in conjunction with some sort of outflow.

Both the Keplerian model and the ADAF model presented here do not satisfy general relativistic inner boundary conditions. For example, since a black hole cannot support shear stress, the torque at the event horizon must be zero (Gammie and Popham 1998). And although the spectral calculations might be general relativistic, the dynamical equations for both models are not. Calculations using pseudo-Newtonian gravity so that $\Phi = -GM/(r - r_s)$ and $\Omega^2 = GM/(r(r - r_s))$ have been done for ADAFs (Narayan *et al* 1997, Chen *et al* 1997) even in a Kerr metric where the black hole is no longer assumed to be non-rotating (Gammie and Popham 1998). Deviations within $\sim 5\, r_s$ are significant but do not appear to change the resulting spectrum substantially.

With (10.93) and the condition that $\Gamma \gg \Lambda$, it is seen that the entropy of the accreting gas in an ADAF increases as it falls inward. Thus, according to the Schwarzschild condition, ADAFs are unstable to convection (Narayan *et al* 2000). Calculations of ADAFs which include hydrothermal convection (Quataert and Gruzinov 2000) show that angular momentum is transported inward, resulting in a flatter density profile ($\rho \propto r^{-1/2}$) (CDAFs, convection dominated accretion flows). However, while some MHD calculations confirm this (Igumenshchev and Narayan 2002), others show that in the presence of turbulent magnetic fields, the net result of convection is to transport angular momentum *outward* (Hawley *et al* 2001). So although it is not certain what the effect of convection is when both resistive and viscous heating are present, it does appear that ADAF models which include convection without a more complete handling of the magnetic field are not self-consistent (Balbus and Hawley 2002).

A fundamental assumption of two-temperature ADAFs is that viscous heating followed by rapid accretion leads to a decoupling of the ions from the electrons. Although it has been shown that the Coulomb coupling itself may be too weak to maintain a thermal equilibrium, other mechanisms may prevent the formation of a two-temperature plasma. For example, Alfvénic turbulence or plasma waves in the presence of a magnetic field in thermal equipartition preferentially heats the electrons (Begelman and Chiueh 1988, Quataert and Gruzinov 1999). If such turbulence is due to collisionless shocks, the resulting heating rate will exceed the Coulomb rate, Γ^{ie}, if $M \gtrsim 3 \times 10^4\, M_\odot$ (Narayan and Yi 1995b). Thus, although ADAFs may be applicable to XRBs, for Sgr A*, where $M = 2.6 \times 10^6\, M_\odot$, it is not at all clear that a stable large-scale two-temperature plasma develops.

10.4 Comment on X-ray emission from Sgr A*

The X-ray emission from Sgr A* is rather weak in quiescence with $L_x \simeq 0.5\ L_\odot$ in the Chandra band of 2–10 keV. However, Sgr A* has been observed to flare at least once for a few hours, reaching a peak luminosity of $\gtrsim 25\ L_\odot$ (Baganoff *et al* 2001a). The flare had a distinctively harder spectrum, with the spectral index α changing from $\simeq 1.2$ to $\simeq 0.3$ ($L_\nu \propto \nu^{-\alpha}$). The spectral index, variability, and flux of both states place significant constraints on Sgr A* accretion models. For example, the quiescent spectral index is probably too soft to be due to thermal bremsstrahlung emission alone (Liu and Melia 2002); it is more likely due to Comptonization of the radio synchrotron photons (SSC). On the other hand, it is not clear whether the flare emission is due to thermal bremsstrahlung (Liu and Melia 2002) or SSC (Falcke and Markoff 2000). In any case, significant SSC only occurs if the peak electron temperature $\gtrsim 10^{10}$ K, a temperature ADAF models do not reach. A possible solution to this is a combination ADAF plus jet model, which reproduces the observed spectral indices and fluxes of both states rather well (Markoff *et al* 2001). Future simultaneous multi-wavelength observations should determine whether thermal bremsstrahlung, which is extended and mostly decoupled from the radio emission, or SSC, which is directly tied to the sub-mm emission, is dominant for Sgr A*.

10.5 Summary

The Galactic black hole has a mass ($\sim 10^6\ M_\odot$) that is in between that of AGNs ($\gtrsim 10^7\ M_\odot$) and XRBs ($\lesssim 10\ M_\odot$), so one naturally thinks of it as a transition type object. Although some nuclear activity is stellar, the term 'low luminosity AGNs' (LLAGNs) has been used to describe galactic nuclei that appear to contain 'baby' supermassive black holes, such as Sgr A*, which are accreting at a rate that is significantly sub-Eddington (Nagar *et al* 2000). However, explaining the spectrum of these objects has proven to be a bit tricky. For example, for other LLAGNs, the radio emission does not appear to be from ADAFs but rather jets (Falcke *et al* 2000, Ulvestad and Ho 2001). In the case of Sgr A*, the magnetic bremmstrahlung emission may be self-absorbed so an ADAF may still be applicable.

Since XRBs, AGNs, and even LLAGNs (Nagar *et al* 2001) generally have jets, it is natural to expect Sgr A* to have one—but it is the apparent absence of such a jet and any accompanying accretion disk which sets Sgr A* apart from other black hole systems. Although there is evidence that Sgr A* is asymmetric on the scale of less than a few AU or $\lesssim 10^2\ r_s$ (Lo *et al* 1999), there is as yet no evidence for any outflow. Perhaps what truly sets Sgr A* apart is its diet: if no clear plane is defined by the wind sources so that on the average there is little angular momentum in the flow, no large accretion disk will form and even a small one may be unstable on fairly short timescales. Yet, since the flow almost certainly has *some* residual angular momentum, some sort of

circularization region probably exists. Where there are disklike structures, there are often jets.

The accretion picture for the Galactic black hole is far from clear. At this point, photon scattering and absorption, relativistic corrections, and magnetic fields are all treated fairly simply; more detailed models might improve on these. Nonetheless, it seems that the true picture contains some combination of large-scale spherical accretion plus a circularization region (which may or may not form a true disk) and perhaps a small-scale jet.

Acknowledgements

References

Baganoff F K *et al* 2001a *Nature* **413** 45
——2001b *Astrophys. J.* submitted *Preprint* astro-ph/0102151
Balbus S A and Hawley J F 1991 *Astrophys. J.* **376** 214
——2002 *Astrophys. J.* **573** 749
Beckert T, Duschl W, Mezger P and Zylka R 1996 *Astron. Astrophys.* **307** 450
Begelman M C and Chiueh T 1988 *Astrophys. J.* **332** 872
Bondi H 1952 *Mon. Not. R. Astron. Soc.* **112** 195
Bondi H and Hoyle F 1944 *Mon. Not. R. Astron. Soc.* **104** 273
Brandenburg A, Nordlund A, Stein R F and Torkelsson U 1995 *Astrophys. J.* **446** 741
Chandrasekhar S 1939 *Introduction to the Theory of Stellar Structure* (New York: Dover)
Chen X, Abramowicz M A and Lasota J 1997 *Astrophys. J.* **476** 61
Coker R F and Markoff S 2001 *IAU Symp.: Galaxies and their Constituents at the Highest Angular Resolution* vol 205, p 43
Coker R and Melia F 1997 *Astrophys. J.* **488** L149
——1999 *ASP Conf. Ser. 186: The Central Parsecs of the Galaxy* ed H Falcke, A Cotera, W Duschl, F Melia and M Rieke (San Francisco: ASP) p 214
——2000 *Astrophys. J.* **534** 723
Coker R, Melia F and Falcke H 1999 *Astrophys. J.* **523** 642
Esin A A, McClintock J E and Narayan R 1997 *Astrophys. J.* **489** 865
Falcke H, Goss W, Matsuo H, Teuben P, Zhao J and Zylka R 1998 *Astrophys. J.* **499** 731
Falcke H and Markoff S 2000 *Astron. Astrophys.* **362** 113
Falcke H, Nagar N M, Wilson A S and Ulvestad J S 2000 *Astrophys. J.* **542** 197
Frank J, King A and Raine D 2002 *Accretion Power in Astrophysics* 3rd edn ed J Frank, A King and D Raine (Cambridge: Cambridge University Press)
Gammie C F and Popham R 1998 *Astrophys. J.* **498** 313
Genzel R, Pichon C, Eckart A, Gerhard O E and Ott T 2000 *Mon. Not. R. Astron. Soc.* **317** 348
Genzel R, Thatte N, Krabbe A, Kroker H and Tacconi-Garman L 1996 *Astrophys. J.* **472** 153

Hawley J F, Balbus S A and Stone J M 2001 *Astrophys. J.* **554** L49

Hawley J F, Gammie C F and Balbus S A 1995 *Astrophys. J.* **440** 742

Hoyle F and Lyttleton R 1939 *Proc. Camb. Phil. Soc.* **35** 592

Ichimaru S 1977 *Astrophys. J.* **214** 840

Igumenshchev I V and Narayan R 2002 *Astrophys. J.* **566** 137

Ipser J R and Price R H 1982 *Astrophys. J.* **255** 654

Liu S and Melia F 2001 *Astrophys. J.* **561** L77

——2002 *Astrophys. J.* **566** L77

Lo K Y, Shen Z, Zhao J-H and Ho P T P 1999 *ASP Conf. Ser. 186: the Central Parsecs of the Galaxy* ed H Falcke, A Cotera, W Duschl, F Melia and M Rieke (San Francisco, CA: ASP) p 72

Mahadevan R 1997 *Astrophys. J.* **477** 585

Markoff S, Falcke H, Yuan F and Biermann P L 2001 *Astron. Astrophys.* **379** L13

Melia F 1988 *J. Comput. Phys.* **74** 488

——1992 *Astrophys. J.* **387** L25

Melia F and Coker R 1999 *Astrophys. J.* **511** 750

Melia F and Kowalenko V 2001 *Mon. Not. R. Astron. Soc.* **327** 1279

Melia F, Liu S and Coker R 2001 *Astrophys. J.* **553** 146

Moncrief V 1980 *Astrophys. J.* **235** 1038

Nagar N M, Falcke H, Wilson A S and Ho L C 2000 *Astrophys. J.* **542** 186

Nagar N M, Wilson A S and Falcke H 2001 *Astrophys. J.* **559** L87

Narayan R, Igumenshchev I V and Abramowicz M A 2000 *Astrophys. J.* **539** 798

Narayan R, Kato S and Honma F 1997 *Astrophys. J.* **476** 49

Narayan R, Mahadevan R and Quataert E 1998 *Theory of Black Hole Accretion Disks* ed M A Abramowicz, G Bjornsson and J E Pringle (Cambridge: Cambridge University Press) p 148

Narayan R and Yi I 1994 *Astrophys. J.* **428** L13

——1995a *Astrophys. J.* **444** 231

——1995b *Astrophys. J.* **452** 710

Pacholczyk A 1970 *Radio Astrophysics* (San Francisco: Freeman)

Parker E N 1960 *Astrophys. J.* **132** 175

Quataert E and Gruzinov A 1999 *Astrophys. J.* **520** 248

——2000 *Astrophys. J.* **539** 809

Rybicki G and Lightman A 1979 *Radiative Processes in Astrophysics* (New York: Wiley)

Shakura N I and Sunyaev R A 1973 *Astron. Astrophys.* **24** 337

Shapiro S 1973 *Astrophys. J.* **180** 531

Shapiro S and Teukolsky S 1983 *Black Holes, White Dwarfs, and Neutron Stars* (New York: Wiley)

Spitzer L 1962 *Physics of Fully Ionized Gases* 2nd edn (New York: Interscience)

Stepney S and Guilbert P W 1983 *Mon. Not. R. Astron. Soc.* **204** 1269

Tamblyn P *et al* 1996 *Astrophys. J.* **456** 206

Ulvestad J S and Ho L C 2001 *Astrophys. J.* **562** L133

Zylka R, Mezger P, Ward-Thomson D, Duschl W and Lesch H 1995 *Astron. Astrophys.* **297** 83

Chapter 11

Radio and X-ray emission from the Galactic Black Hole

Heino Falcke
Max-Planck-Institut für Radioastronomie, Bonn, Germany

Finally, the radio properties of the Galactic Center black hole (Sgr A*) are reviewed: variability, size and position, spectrum, and polarization. Radio and X-ray emission are discussed within the framework of black hole plasma jet models and simple equations for the emission are derived. It is also shown that the radio emission can be used to actually image the event horizon of the black hole in the near future.

11.1 Introduction

We have already seen in chapter 8 that the evidence for the presence of a dark mass in the Galactic Center is very strong. A central point mass of about $3 \times 10^6 \ M_\odot$ seems to coincide with the compact radio source Sgr A* (figure 11.1). The existence of Sgr A* has always been considered a good sign for a black hole itself. In fact, based on analogous detections of compact radio cores in the nuclei of active galaxies, the existence of Sgr A* was predicted by Lynden-Bell and Rees (1971). Balick and Brown (1974) detected the source in one of the early VLBI (Very Long Baseline Interferometry) experiments and a couple of years later named it Sgr A* (by simply adding an asterisk to the name of the nuclear radio region Sgr A—see Palmer and Goss (1996) for an account of the history of naming sources in the Galactic Center).

Ever since, Sgr A* has been the focus of great attention. From the near-infrared speckle observations (chapter 8) we know that it is at the very center of the gravitational potential. However, until today one remains unsure about the exact nature of the detected emission in radio and X-ray bands. However, understanding the radiation spectrum of Sgr A* will have important implications
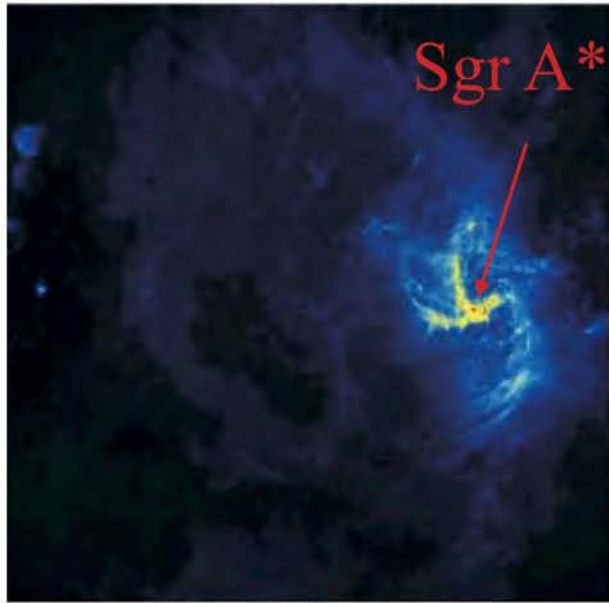
Figure 11.1. VLA radio map of Sgr A West, a spiral-like pattern of thermal ionized gas that appears to be falling into the very center of the galaxy. Near its center is Sgr A*, a pointlike radio source that many suspect is the nucleus of the Milky Way and indicates the presence of a black hole. Figure courtesy of K Y Lo. See also color section.

for all other supermassive black holes, since most galaxies will have black holes that are as inactive as the Galactic Center and we are investigating the low-power end on the black hole activity scale. In fact, surveys have shown that a large number of nearby galaxies host radio sources very similar to Sgr A* in our Milky Way (Wrobel and Heeschen 1984, Nagar *et al* 2000, Falcke *et al* 2000). Hence, what we learn about Sgr A* is quite typical for the (rather silent) majority of black holes in the universe.

In the following section we will attempt to summarize the main radio properties of the 'Galactic Black Hole'. In the subsequent sections we will try to explain the main emission characteristics of Sgr A* in an almost back-of-the-envelope fashion. Finally we will look ahead to how this source could finally make the theoretical concept of an event horizon observable in the near future. A comprehensive review of this can be found in Melia and Falcke (2001).

11.2   Radio properties of Sgr A*

Most of the direct information about Sgr A* is available at radio wavelengths. This includes the total intensity spectrum, the variability, the polarization, and the

Figure 11.2. The Green Bank Interferometer (GBI; left) which was instrumental in detecting Sgr A* and the Very Large Array (VLA; right). Both interferometers are operated by the National Radio Astronomy Observatory (NRAO) in the United States and were extensively used to monitor the variability of Sgr A*.

source structure. In the following we will go through these various issues step by step.

11.2.1 Variability of Sgr A*

The first evidence for a a very compact structure in a radio source often comes from its variability. And indeed soon after its discovery, Sgr A* was established as a variable source (Brown and Lo 1982) and extended campaigns were set up to monitor this variability. The most extensive data sets were obtained using the Very Large Array (VLA) and the Green Bank Interferometer (GBI; see figure 11.2). The most recent data from these instruments are presented in Zhao *et al* (2001) and Falcke (1999). The amplitude of variability can reach up to 200% for strong flares. The degree of variability seems to increase with increasing frequency. Strong flares seem to occur on timescales of 100 days (see figure 11.3). Zhao *et al* (2001) even claim to have found a periodicity of 106 days on which such strong flares occur regularly. The shortest timescale of radio variability was probably found by Bower *et al* (2002a) at 15 GHz: 20% within 1 h.

Some of the radio variability may be due to scintillation due to a foreground screen (see p 315), but at least the large-amplitude flares at higher radio frequencies are most likely to be intrinsic. The fastest variations are fundamentally limited by the source size. We can then convert the measured timescale τ and amplitude $\Delta S/S$ of the variability into a limit on the characteristic size R of Sgr A*. For a given maximum signal speed $v_{max} \lesssim c$ in the emitting plasma, one obtains

$$R < (\Delta S/S)^{-1}\tau \times v_{max}. \tag{11.1}$$

For $(\Delta S/S) = 20\%$ and $\tau = 1$ hr we get $R < 5 \times 10^{14}$ cm $\times (v_{max}/c)$, which is 36 Astronomical Units (AU) or 4.5 milli-arcseconds (mas) at the Galactic
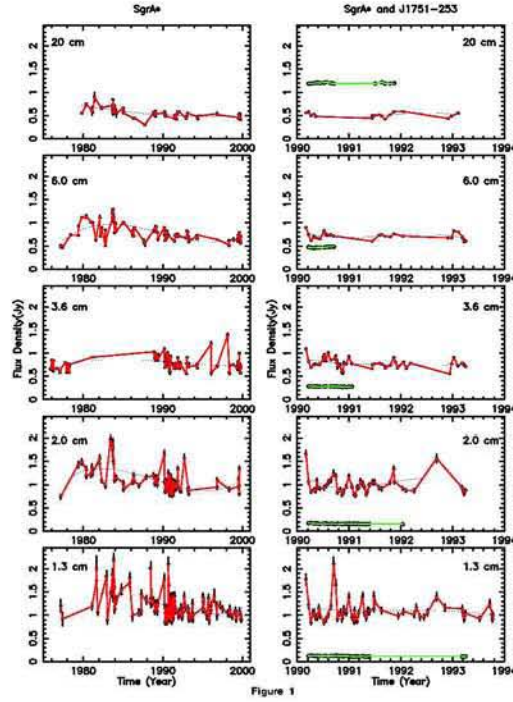
Figure 11.3. Variability of Sgr A* as observed with the VLA at different wavelengths (given in the top left-hand corner of each frame). From Zhao *et al* (2001).

Center distance of 8 kpc. Based on the variability timescale alone the radio emission therefore has to come from a region smaller than a planetary system. In fact, as we will learn in the next section, direct radio imaging shows that the emission comes from an even smaller scale.

Alternatively, taking the much longer 100 day timescale we find a characteristic size of $R < 2.7 \times 10^{17}$ cm $\times (v_{max}/c)$ or 2.3 arcseconds. This is relatively large and does not appear to be a useful limit for the size of Sgr A*. It is more likely that this variability signals some other underlying physical process, e.g. a process related to the accreting material powering the source.

If the accreting gas is in orbit around the central point mass $M_\bullet$ the characteristic velocity will be set by Keplerian velocity[1] $v = \sqrt{GM_\bullet/R}$ and the corresponding timescale is set by $\tau = 2\pi R/v$, yielding

$$R \approx \left(\frac{\tau}{2\pi}\right)^{2/3} (GM_\bullet)^{1/3} = 0.9 \times 10^{15} \text{ cm} \left(\frac{\tau}{106 \text{ days}}\right)^{2/3} \left(\frac{M_\bullet}{3 \times 10^6 M_\odot}\right)^{1/3}.$$

(11.2)

[1] Note that the sound speed of infalling gas in optically thin accretion models is typically of similar order and hence the same numbers apply.

This corresponds to about 1000 Schwarzschild radii or 8 mas. The accretion radius (chapter 10) of the hot X-ray gas in the Galactic Center is somewhat further out, but it is not impossible that instabilities in the capture process of surrounding gas by the black hole are the ultimate source of the slow, large-amplitude variability. If the flares are indeed periodic as claimed by Zhao *et al* (2001), then one probably has to think of a star orbiting Sgr A* at this distance which modulates the accretion flow.

However, in stellar mass black holes, quasi-periodic signals are often related to beat and precession frequencies of accretion disks close to the black hole. If true, this can be used to derive information about the black hole spin. Such a scenario has been considered by Liu and Melia (2002). Based on such a model and the 100 day periodicity, they derive a black hole spin parameter (e.g. section 11.4) of Sgr A* of $a \simeq 0.1$. It is difficult to assess which of these interpretations is correct.

The fastest variations are expected at the shortest wavelengths, since high-frequency emission—in essentially all models for Sgr A*—is typically produced at the smallest scales. The problem with high-frequency measurements of Sgr A* is that these observations are extremely sensitive to weather and the low elevations of the source encountered in typical observations from the northern hemisphere. Because of the large confusing flux in the Galactic Center, interferometers have to be used. The flux density measured on an interferometer baseline can be artificially reduced by rapid changes in the atmospheric opacity and loss of coherence (of instrumental or atmospheric origin) which are difficult to track. Early observations (Wright and Backer 1993, Zylka *et al* 1995) already indicated possibly strong variations in the sub-millimeter wavelength region. This was later strengthened by Tsuboi *et al* (1999) and very recently Zhao *et al* (2001) claim relatively strong sub-millimeter flares in Sgr A* from measurements with the sub-millimeter array (SMA) in Hawaii. These results will certainly become increasingly important in the years to come since, as we will see later, this emission most certainly comes directly from the immediate environment of the event horizon.

11.2.2   Size of Sgr A*—VLBI observations

A major issue for a long time has been the exact size and structure of Sgr A*. In extragalactic sources the VLBI technique has allowed one to probe deep into the hearts of active galactic nuclei (AGNs) and to resolve the radio structures into relativistically outflowing plasma jets (see, e.g., Zensus 1997). In a VLBI experiment radio telescopes distributed over a continent or the entire world are synchronized by atomic clocks and jointly observe one source. The incoming radio waves are digitized and stored (usually on magnetic tapes). Later, the digitized waves are correlated in a specialized computer recreating a virtual interferometer. Such a virtual interferometer will have a spatial resolution similar to that of a giant telescope with the diameter of roughly the separation of the
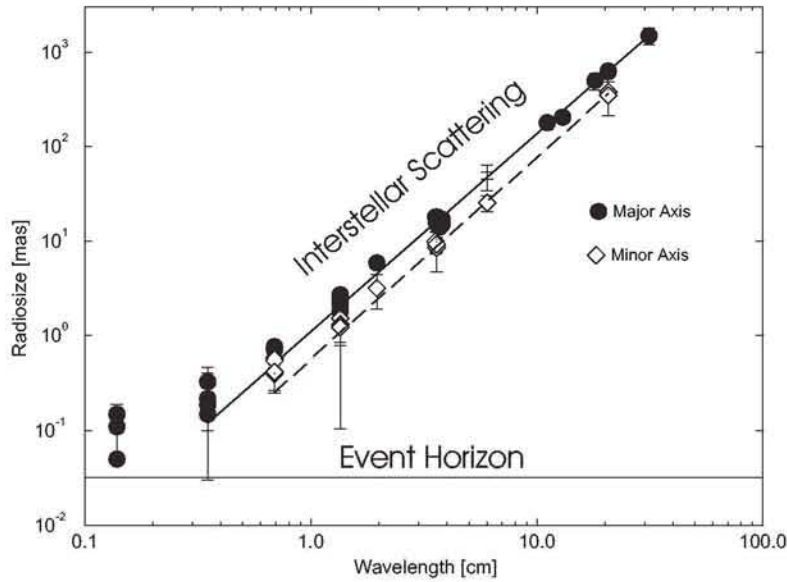
Figure 11.4. The major source axis (filled circles) of Sgr A* and the minor source axis (open diamonds) as measured by VLBI plotted *versus* wavelength (adapted from Krichbaum *et al* 1999). The inclined lines show the λ^2 scattering law and the horizontal line shows the size scale expected for the visual imprint of the event horizon (Falcke *et al* 2000).

individual telescopes (called 'baselines'). This technique has yielded by far the highest-resolution images in astronomy (i.e. <50 μarcseconds at 86 GHz). The image quality improves with the number of participating telescopes and the two major VLBI arrays today are the European VLBI Network (EVN) and the Very Long Baseline Array (VLBA) in the United States.

Of course, a prominent source such as Sgr A* has been a prime target for VLBI experiments over the last 25 years. It is from these observations that we have direct information on the size of the source—or at least have very tight upper limits.

It was, however, quickly realized that, despite its relative proximity, detecting the true structure of Sgr A* is unusually difficult compared to other galactic nuclei. The reason is that interstellar material in the line-of-sight towards the Galactic Center scatter broadens the VLBI image. This produces a characteristic λ^2 law (e.g. Scheuer 1968) for the size of Sgr A* as a function of the wavelength λ (Davies *et al* 1976, van Langevelde *et al* 1992, Yusef-Zadeh *et al* 1994, Lo *et al* 1998). The scattering size apparently has not changed over a decade (Marcaide *et*
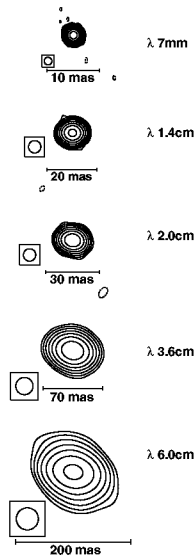
Figure 11.5. Contour plots of VLBA images of Sgr A* at wavelengths of λ = 6.0, 3.6, 2.0, 1.35, and 0.7 cm. These images are smoothed to a circular beam of FWHM = $2.62\lambda_{cm}^{1.5}$ mas as shown at the bottom left-hand corner on each image. The contours are 2 mJy beam^{-1} $\times$ (−2, 2, 4, 8, 16, 32, 64, 128, 256). Figure from Lo *et al* (1999).

al 1999). For that reason the *intrinsic* size and structure of Sgr A* has remained obscure until today (figures 11.4 and 11.5). The measured size of Sgr A* is given by (Lo *et al* 1998)

$$\theta_{minor} = 0.76 \, \text{mas}(\lambda/\text{cm})^2 \qquad \theta_{major} = 1.42 \, \text{mas}(\lambda/\text{cm})^2. \qquad (11.3)$$

The front line of this research is currently at wavelengths of 7 mm and shorter. Lo *et al* (1998) and Krichbaum *et al* (1993, 1998) claim to have seen evidence for a deviation from the scattering law at these wavelengths. Other experiments did not directly confirm this (Rogers *et al* 1994, Bower and Backer 1998, Doeleman *et al* 2001). The main problem is that Sgr A* is never above 25° elevation at the VLBA and mm-VLBI is strongly affected by the variable atmosphere. Significant distortions of the phase (of the radio waves) can happen on timescales of 10 s at the short wavelengths. Therefore mm-VLBI observations of Sgr A* are difficult to calibrate and are always subject to intense scrutiny and nagging doubts.

However, the rapid variability of Sgr A* with timescales of 10 minutes (X-rays) to 3 h (15 GHz radio) suggests that time-variable structure may exist on similarly short timescales. We can estimate a scale for adiabatic cooling of a

plasma by using equation (11.3), convert to a linear size at the Galactic Center distance $D_{GC} = 8$ kpc, and divide by the maximal sound speed of a relativistic plasma (photon gas; see, e.g., Königl 1980) $c_s = c/\sqrt{3}$. We get

$$t_{cool} \gtrsim \theta_{major} D_{GC}/c_s = 1.3 \, h \left(\frac{\lambda}{7 \, mm} \right)^2 . \tag{11.4}$$

Hence, if for example, matter is ejected during one of the big X-ray flares near the black hole it could, in principle, cool and fade away within a few hours at most radio wavelengths[2]. Consequently, there is no reason why VLBI observations separated by one day will always look the same.

11.2.3 Position of Sgr A*

The exact location of Sgr A* is a very important factor for its interpretation. This allows one to investigate whether the source is indeed in the Galactic Center (and not behind) and to set a lower limit on its mass. The position can be determined relatively accurately with radio observations.

This requires the use of so-called 'phase-referencing' observations. In this type of experiment the telescopes of a VLBI array are switched rapidly from one source to another, where both sources should be within one isoplanatic patch of the atmosphere (typically $1°–5°$). Within this patch the radiation passes roughly through the same atmospheric irregularities. The switching also has to happen within the coherence time of the atmosphere—at mm wavelengths the telescopes switch sources every 10 s. The phase difference of the incoming wavefronts between the two observed sources can then be used to determine the relative positions of the sources. By fixing a grid of phase-referenced sources—radio quasars at cosmological distances—one can then establish an absolute coordinate system, called the International Celestial Reference System (ICRF; see Ma *et al* 1998). Since 1997 the ICRF has been the fundamental reference system for astronomy as adopted by the International Astronomical Union (IAU).

Attempts have been made to relate the *absolute* position of Sgr A* to bright background quasars. The averaging of VLA observations by Yusef-Zadeh *et al* (1999) yielded a position for Sgr A* at the epoch 1992.4 of

$$\alpha(1950) = 17^h 42^m 29.3076^s \pm 0.0007^s, \delta(1950) = -28°59'18.484'' \pm 0.014''$$
$$\tag{11.5}$$
$$\alpha(2000) = 17^h 45^m 40.0383^s \pm 0.0007^s, \delta(2000) = -29°00'2^s.069'' \pm 0.014''.$$
$$\tag{11.6}$$

A position using VLBA observations was also derived by Rogers *et al* (1994), which agrees with this position within $0.2''$. In yet another experiment

---

[2] Note that equation (11.4) is based solely on the *observed* upper limit on the source size. Most models suggest that the intrinsic size of Sgr A* grows less rapidly than λ^2 (e.g. linearly) and hence the cooling timescale would grow accordingly more slowly.
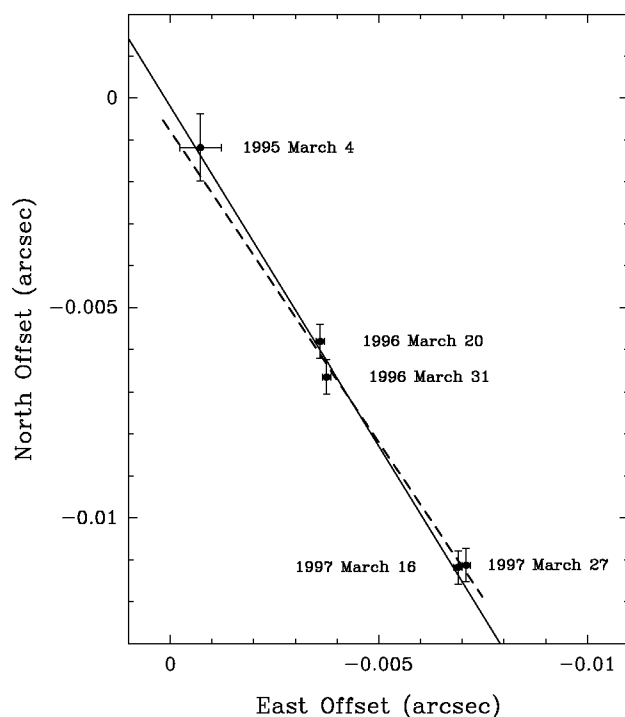
Figure 11.6. Position of Sgr A* relative to a background quasar (J1745–283) on the plane of the sky determined from VLBA observations. North is to the top and East to the left. Each measurement is indicated with the date of observation and 1 − σ error bars. The broken line is the best-fit proper motion, and the full line gives the orientation of the Galactic plane. Figure from Reid *et al* (1999).

Menten *et al* (1997) were able to relate the position of Sgr A* to the position of the near-infrared stars (emitting a radio line) surrounding it. This allows one to locate Sgr A* in a near-infrared image and try to find a counterpart.

Finally, one can also measure the *relative* position of Sgr A* with respect to faint background quasars which are much closer than the ICRF reference sources. Assuming that these sources are without motion on the sky (because of their cosmological distances), Reid *et al* (1999) (figure 11.6; see also Backer and Sramek 1999) find a proper motion for Sgr A* of −3.33 ± 0.1 mas yr^{-1} (E) and −4.94 ± 0.4 mas yr^{-1} (N), corresponding to −5.90 ± 0.35 mas yr^{-1} and +0.20 ± 0.30 mas yr^{-1} in Galactic longitude and latitude. This agrees very well with the apparent motion expected for a source at the Galactic Center due to the *Galactic rotation of the solar system* (220 km s^{-1}). This implies that Sgr A* is indeed at the center of the Galaxy and has very little motion of its own ($v_{\text{Sgr A*}} < 15$ km s^{-1}). This is interesting, since in chapter 8 it was

shown that stars near Sgr A* move in the deep potential well with velocities up to 1500 km s^{-1}. The most likely reason for this is, of course, that Sgr A* itself causes this potential well. One can use the assumption of equipartition of momentum between the fastest stars ($m_* v_*$) and Sgr A* ($M_{\text{Sgr A}*}\ v_{\text{Sgr A}*}$) to infer a lower limit on the mass of Sgr A* from the VLBI proper motion studies:

$$M_{\text{Sgr A}*} \gtrsim 1000\ M_\odot \left(\frac{m_*}{10\ M_\odot}\right)\left(\frac{v_*}{1500\ \text{km s}^{-1}}\right)\left(\frac{v_{\text{Sgr A}*}}{15\ \text{km s}^{-1}}\right)^{-1}. \qquad (11.7)$$

Numerical modeling of n-body interactions suggests that under most conditions this lower limit can be as high as $10^5\ M_\odot$ (Reid *et al* 1999). This mass is way beyond those of any stellar object and hence it is very reasonable to assume that essentially all the dark mass of 2–3 × $10^6\ M_\odot$ is concentrated inside Sgr A*—the radio source.

11.2.4 Radio spectrum of Sgr A*

A major input for modeling the nature of the central black hole candidate of our Galaxy is the emission spectrum. Typical luminous black holes, e.g. those in quasars, emit over a broad range in frequencies. In contrast, for Sgr A* only radio emission was reliably detected over many years, with the late addition of X-ray emission. One reason for this dimness is certainly the low accretion rate and quite plausibly the presence of a radiatively deficient accretion flow (chapter 10).

At frequencies below 1 GHz the spectrum is essentially undetermined. First of all, the scattering size of Sgr A* becomes larger than 1 arcsecond and the source starts to blend with its surroundings. Second, the Sgr A complex becomes optically thick at low frequencies (Pedlar *et al* 1989) and Sgr A* may be obscured. It is also possible, but less likely, that the claimed low-frequency turnover in the spectrum (Davies *et al* 1976) has an intrinsic nature.

At higher frequencies the radio spectrum of Sgr A* has been measured with great accuracy in various campaigns up to the THz regime (e.g. Wright and Backer 1993, Zylka *et al* 1995, Serabyn *et al* 1997, Falcke *et al* 1998). Since the source is variable (see figure 11.3), it is useful to consider either simultaneous or time-averaged spectra. Such an average spectrum is shown in figure 11.7 compiled by Melia and Falcke (2001). The overall radio spectrum is slightly inverted, i.e. it has a positive spectral index ($\alpha \simeq 0.2$, $S_\nu \propto \nu^\alpha$) in the GHz regime. The average radio flux density is around 1 Jy. At the highest radio frequency ($\gtrsim 100$ GHz), the spectrum seems to become even more inverted until it abruptly cuts off somewhere in the far-infrared. This upturn and subsequent cut-off has been interpreted as an effect of the finite size of the central object with a size scale as expected for a 3 × $10^6\ M_\odot$ black hole (e.g. Falcke and Biermann 1994 and later). The up-turn in the spectrum at sub-mm wavelengths is often referred to as the 'sub-mm bump'.
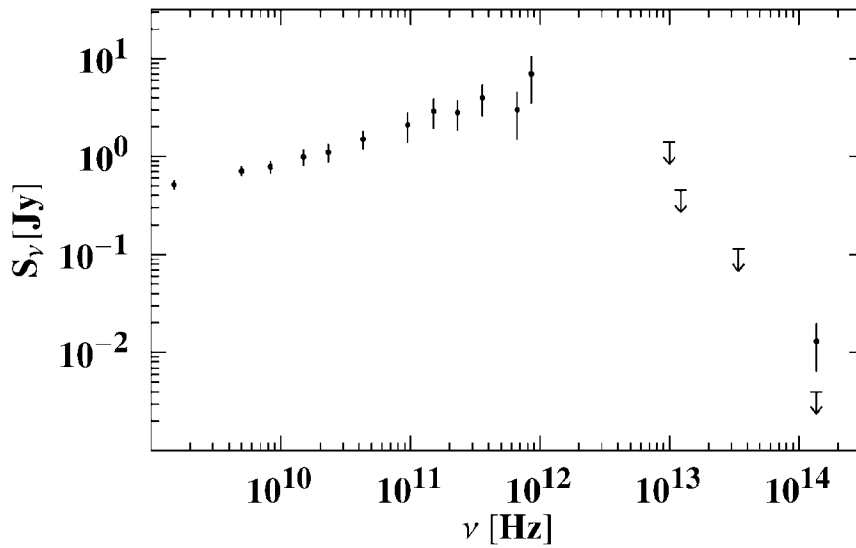
Figure 11.7. Time-averaged spectrum—flux density *versus* frequency—of Sgr A* from radio to the near-infrared as compiled by Melia and Falcke (2001). The error bars in the radio regime indicate variability (one standard deviation).

11.2.5 Polarization of Sgr A*

Finally, as a relatively recent development, the polarization properties of Sgr A* are now relatively well established. For a long time it was generally thought that, in marked contrast to more luminous AGNs, Sgr A* was unpolarized. This is true for linear polarization in the GHz radio regime. Bower *et al* (1999a) found an upper limit of $\leq 0.1\%$ to the linear polarization at cm waves; however, they also found plenty of circular polarization.

The results for linear polarization at low frequencies were obtained with a rarely used technique in radio astronomy, called spectro-polarimetry. This allows one to look for polarization in small frequency bands. Usually in continuum observations one averages the polarization of the radiation over the available bandwidth δv. However, since Sgr A* may be embedded in a dense plasma, Faraday rotation in the accretion flow or a foreground Galactic screen could lead to a rotation of the linear polarization vector even within the small bandwidth.

Faraday rotation is produced when radio waves pass through an ionized and magnetized medium. Since left and right circularly polarized waves have different refractive indices for a given magnetic field orientation, a wavelength-dependent delay is induced that rotates the position angle ϕ of the linear polarization vector by an amount

$$\Delta\phi = \text{RM}\lambda^2. \tag{11.8}$$

The parameter RM is called the rotation measure and can be determined by measuring the position angle of the linear polarization vector ϕ at different wavelengths. For a given frequency bandwidth $\delta \nu$, significant de-polarization is obtained if $\Delta \phi \sim 1$ rad. Hence, for a typical VLA bandwidth of $\Delta \nu = 50$ MHz at 4.8 GHz a rotation measure of RM $= 10^4$ rad m^{-2} of any foreground material would destroy any intrinsic polarization signal. Such a value for RM is large compared to what is seen in other AGNs but cannot be excluded in the Galactic Center. By Fourier-transforming spectro-polarimetric data (to look for periodic signals due to the fast rotation of the polarization vector as a function of frequency), Bower *et al* (1999a) were able to set the 0.1% limit and also exclude rotation measures below RM $\lesssim 10^7$ rad m^{-2} for a homogeneous foreground screen as the cause for the low polarization.

Later, Aitken *et al* (2000) made linear polarization observations with a single dish sub-millimeter wave telescope and found $\sim 10\%$ linear polarization above 150 GHz. This was confirmed with an interferometer by Bower *et al* (2002b) and they also found evidence for a large rotation measure $\lesssim 10^6$ rad m^{-2} plus some evidence for intrinsic depolarization towards lower frequencies. This may be the first direct evidence for a hot accretion flow around Sgr A*. At the moment of writing this is a strongly developing field which promises many new insights in the future. For example, one can use the measured RM to limit the accreting material engulfing Sgr A* (Agol 2000) (see also p 335).

As a big surprise, Bower *et al* (1999b) also found strong circular polarization at the 0.3–1% level. This is unusual because typical AGN polarization is more linear than circular and it can be used to constrain the electron content and distribution in Sgr A* (see section 11.3.4). Interestingly, the circular polarization itself turned out to be variable. At higher frequencies the variability as well as the fractional polarization increases (Bower *et al* 2002a). Figure 11.8 summarizes the currently known polarization properties of the Sgr A* spectrum.

11.3 Radio and X-ray emission from a black hole jet

A very common feature of active black holes in the radio regime is the compact radio core with its characteristic, flat spectrum. In luminous quasars the cores have been known for many years. Studying these radio cores with VLBI has allowed us to make the most detailed images of the physics and environment of black holes (Zensus 1997). Such flat-spectrum radio cores have also been found in many nearby galaxies which show signs of nuclear activity (Wrobel and Heeschen 1984, Nagar *et al* 2000, Falcke *et al* 2000). A well studied example is M81*, the compact core in the nearby galaxy M81, which shares many characteristics with Sgr A* (Reuter and Lesch 1996, Bietenholz *et al* 2000, Brunthaler *et al* 2001). In essentially all cases these cores are related to relativistic outflows or jets. For this reason, we start with the simplest assumption, namely that Sgr A* is not very different from these, and we will discuss in the following how to obtain the
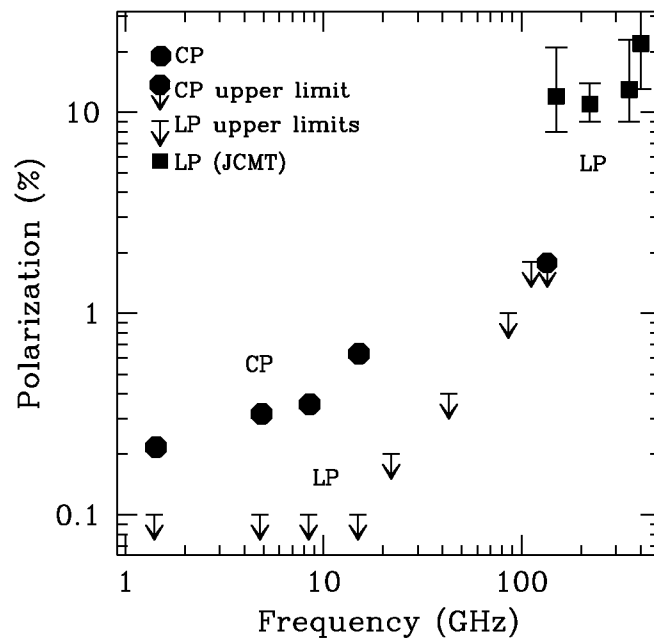
Figure 11.8. The average fractional circular polarization of Sgr A* and upper limits to the linear polarization from Bower *et al* (2002; and references therein). In the top right-hand corner we show the linear polarization values given in Aitken *et al* (2000) from single-dish values. The error bars are 1σ errors.

observed radio spectra within the context of jet physics.

11.3.1   The flat radio spectrum

The fact that flat radio spectra for radio cores are so ubiquitous suggests that this is a very robust feature that must arise naturally. This is indeed the case for initially collimated, then freely expanding, supersonic radio plasmas. Why is this so?

Let us consider a plasma jet ejected from the vicinity of a black hole. Mechanisms for this collimated launching of jets have been discussed in the literature (see, e.g., Ferrari 1998 and references therein) and are mostly magneto-hydrodynamic (MHD) in origin. Observationally, jets span an enormous range of spatial scales—from milliparsecs to megaparsecs—and maintain their basic structure for long stretches (see, e.g., Bridle and Perley 1984, figure 2). Here, we consider the part where the jet has left the acceleration and collimation region and is essentially in a free expansion. If the jet has not yet propagated and expanded too far, it is usually a good assumption to assume that the jet is highly over-pressured with respect to the external medium. We use a cylindrical coordinate system where z is along the jet axis and r is perpendicular to it.

Let us assume the jet plasma moves in the forward direction with a relativistic and almost constant proper velocity (bulk speed)

$$v_z = \gamma_j \beta_j c \qquad (11.9)$$

along the jet axis. The sideways expansion will happen with the respective sound velocity

$$v_s = \gamma_s \beta_s c \qquad (11.10)$$

if we can ignore the external pressure and we are well beyond the sonic point where we can neglect adiabatic losses.

Here we use the well-known definition of the relativistic Lorentz factor and the dimensionless velocity,

$$\gamma = \sqrt{\frac{1}{1 - \beta^2}} \qquad \text{and} \qquad \beta = \frac{v}{c}. \qquad (11.11)$$

With longitudinal and lateral expansion having constant velocity the plasma will expand into a cone with a half opening angle

$$\phi \simeq \frac{1}{\mathcal{M}} \qquad \mathcal{M} = \frac{\gamma_j \beta_j}{\gamma_s \beta_s} \qquad (11.12)$$

where $\mathcal{M}$ is the relativistic Mach number (see Königl 1980). The shape is given by

$$r = \frac{z}{\mathcal{M}}. \qquad (11.13)$$

This naturally resulting conical structure is the basis for the self-similar structure of jets.

The scaling of the relevant parameters for calculating the synchrotron emission, electron density n_e, and magnetic field B can be obtained from simple conservation laws. First, we demand that the particle number N_e is conserved along the flow and set the total mass flux to

$$\dot{M}_j = \frac{d(m_p N_e)}{dt} = \rho v_z A = \text{constant}. \qquad (11.14)$$

Note that the total mass flux is determined by the protons in the fully ionized plasma and we assume charge balance between electrons and protons ($N_e = N_p$); ρ is the mass density and $A = \pi r^2$ the cross section of the jet. The particle density n_e is then given by

$$n_e(r) = \frac{\dot{M}_j}{m_p \times \gamma_j \beta_j c \times \pi r^2} \propto r^{-2}. \qquad (11.15)$$

We can use the same argument to get the scaling for the magnetic field, by demanding that the comoving magnetic field energy in a turbulent plasma is conserved:

$$\dot{E}_B = L_B = \rho_B v_z A = \text{constant}. \qquad (11.16)$$

Here we use L_B as a measure for the magnetic power fed into the jet. The energy density of the magnetic field is given by

$$\rho_B = B_j^2/8\pi \qquad (11.17)$$

and consequently we get

$$B_j(r) = \sqrt{\frac{8L_B}{\gamma_j\beta_j c \times r^2}} = 36 \text{ G } (\gamma_j\beta_j)^{-1/2} \left(\frac{L_B}{1000L_\odot}\right)^{1/2} \left(\frac{r}{R_s}\right)^{-1} \propto r^{-1} \qquad (11.18)$$

where for the Galactic Center we have a Schwarzschild radius of

$$R_S = \frac{2GM_\bullet}{c^2} = 0.9 \times 10^{12} \text{ cm} \left(\frac{M_\bullet}{3 \times 10^6 M_\odot}\right). \qquad (11.19)$$

Of course, this implies that the energy content in a magnetic field and relativistic particles retains a fixed ratio throughout the jet. One therefore relates these two crucial parameters of a jet by an 'equipartition relation' such that the total energy in the particles is a fraction k of the total energy in the magnetic field. For simplicity we assume that all electrons are of the same energy $E_e = \gamma_e m_e c^2$, with γ_e being the electron Lorentz factor characterizing the internal energy or temperature of the electrons (not to be confused with the bulk Lorentz factor of the entire flow)[3]. We can equate the energy densities,

$$n_e\gamma_e m_e c^2 = k\frac{B^2}{8\pi} \qquad \text{and yield } n_e = \frac{kB^2}{8\pi\gamma_e m_e c^2}. \qquad (11.20)$$

Here we only consider the internal energy of the jet. The total energy of the jet will of course be still dominated by the kinetic energy of the protons—but not by a huge factor. A proper discussion of the *total* energy budget requires solution of the relativistic Bernoulli equation and is discussed in Falcke and Biermann (1995).

To calculate the radio emission, we need to know the scaling of the synchrotron emission and absorption coefficients. This can be obtained, e.g. from Pacholczyk (1970; equations (3.43) and (3.44)) for electrons with a pitch angle α_e with respect to the magnetic field. The emission and absorption coefficients are, respectively,

$$\epsilon_\nu = n_e\frac{\sqrt{3}e^3}{4\pi m_e c^2}B \sin\alpha_e F\left(\frac{\nu}{\nu_c}\right) \qquad (11.21)$$

and

$$\alpha_\nu = n_e c^2\sqrt{\frac{3e}{4\pi m_e^3 c^5}}\frac{\sqrt{3}e^3}{4\pi m_e c^2}(B \sin\alpha_e)^{3/2}\nu_c^{-5/2}K_{5/3}\left(\frac{\nu}{\nu_c}\right) \qquad (11.22)$$

[3] The results will not be very different for a thermal distribution of electrons or a power-law distribution with a low-energy cut-off around γ_e.

with

$$v_c = \frac{3\gamma_e^2 e}{4\pi m_e c} B \sin \alpha_e. \tag{11.23}$$

$F(x)$ is a function with asymptotic limit

$$F(x) \sim \frac{4\pi}{\sqrt{3}\Gamma(1/3)} \left(\frac{x}{2}\right)^{1/3} \qquad \text{for } x \ll 1$$

which has a maximum at

$$v_{\max} = 0.29 v_c. \tag{11.24}$$

$K_{5/3}(x)$ is the Bessel K function which can be Taylor expanded into $K_{5/3}(x) = 1.43 x^{-5/3}$ for $x \ll 1$. For the pitch angle we can take an average value:

$$\langle \alpha_e \rangle = \arcsin\left(\frac{\int_0^{\pi/2} \sin \alpha \sin \alpha \, d\alpha}{\int_0^{\pi/2} \sin \alpha \, d\alpha}\right) \simeq 52°. \tag{11.25}$$

Using the asymptotic behavior and average pitch angle, we can evaluate the emission and absorption coefficients and obtain handy approximate formulae:

$$\epsilon_\nu = 6.0 \times 10^{-20} \text{ erg s}^{-1} \text{ Hz}^{-1} \text{ cm}^{-3} k \left(\frac{B}{\text{Gauss}}\right)^{8/3} \left(\frac{\gamma_e}{100}\right)^{-5/3} \left(\frac{\nu}{\text{GHz}}\right)^{1/3} \tag{11.26}$$

and

$$\alpha_\nu = 3.5 \times 10^{-14} \text{ cm}^{-1} k \left(\frac{B}{\text{Gauss}}\right)^{8/3} \left(\frac{\gamma_e}{100}\right)^{-8/3} \left(\frac{\nu}{\text{GHz}}\right)^{-5/3}. \tag{11.27}$$

The synchrotron spectrum of a mono-energetic electron distribution will have a characteristic shape consisting of three parts:

(1) an optically thick spectrum with $S_\nu \propto \nu^2$ at frequencies $\nu \ll \nu_{ssa}$ below the self-absorption frequency,
(2) an optically thin spectrum with $S_\nu \propto \nu^{1/3}$ at intermediate frequencies $\nu_{ssa} < \nu < \nu_c$, and
(3) an exponential high-frequency cut-off beyond $\nu \gg \nu_c$.

In most realistic cases for jets the intermediate region will not assume the $\nu^{1/3}$ law, since ν_c and ν_{ssa} are close together leading to a curved spectrum.

The self-absorption frequency can be calculated from equations (11.27) and (11.18), by requiring that the optical depth through the jet, seen under an angle of $\theta \gg \mathcal{M}^{-1}$ from the jet axis, is unity:

$$\tau \arcsin \theta R \alpha_\nu = 1. \tag{11.28}$$

We find

$$\nu_{ssa} = 2.3 \text{ GHz } k^{3/5} \sin^{-3/5}\theta (\gamma_j \beta_j)^{-4/5} \left(\frac{\gamma_e}{100}\right)^{-8/5} \left(\frac{L_B}{1000 L_\odot}\right)^{4/5} \left(\frac{R}{AU}\right)^{-1}.$$

(11.29)

The maximum flux of synchrotron emission is found at a frequency of $\nu_{max} = 0.29\nu_c$ (equation (11.24); see also Rybicki and Lightman 1979, figure 6.6). Using equations (11.23) and (11.18) we find

$$\nu_{max} = 21 \text{ GHz}(\gamma_j \beta_j)^{-1/2} \left(\frac{\gamma_e}{100}\right)^2 \left(\frac{L_B}{1000 L_\odot}\right)^{1/2} \left(\frac{R}{AU}\right)^{-1}.$$

(11.30)

As one can see, as hinted at earlier, both frequencies are within an order of magnitude for near-equipartition situations and both scale with $R^{-1} = (Z/\mathcal{M})^{-1}$ (equation (11.13)).

Since the synchrotron spectrum peaks near these frequencies one also sees that for a given observing frequency the maximum of the emission in the spatial domain will be at one characteristic zone in the jet. A different observing frequency will reveal a different maximum. This effect produces a core shift that is well known in quasar jets. Since the size of the emitting region, $\Delta R \simeq Z/\mathcal{M}$, will be proportional to the distance one also expects a different core size for different frequencies:

$$Z_{max} = 21 \text{ AU } \mathcal{M}(\gamma_j \beta_j)^{-1/2} \left(\frac{\gamma_e}{100}\right)^2 \left(\frac{L_B}{1000 L_\odot}\right)^{1/2} \left(\frac{\nu}{GHz}\right)^{-1}.$$

(11.31)

The effect of a roughly ν^{-1} core size was nicely demonstrated by Bietenholz *et al* (2000) for M81*. For Sgr A* this effect is not visible at cm waves due to the frequency dependence of the scatter broadening.

At the Galactic Center distance of $D = 8$ kpc, 1 AU corresponds to 0.125 mas. For comparison, the claimed size for Sgr A* at 43 GHz by Lo *et al* (1998) was $\lesssim 0.7$ mas. Equation (11.31) predicts a shift of order 0.25 mas $\times$ $(\mathcal{M}/4)(L_B/1000 L_\odot)$ at 43 GHz.

Another useful comparison is the maximum frequency of the entire spectrum. For the smallest size of the system, one Schwarzschild radius $R_S = 0.9 \times 10^{12}$ cm $= 0.06$ AU, we find a maximum frequency of ~ 350 GHz from equation (11.30). Hence it is immediately understandable why the Sgr A* spectrum has to turn over at higher frequencies, beyond the sub-mm bump. In this respect the location of the sub-mm bump is a rough indicator of the size of the black hole in the Galactic Center.

Finally, we can calculate the total spectrum. We know that each frequency is dominated by a relatively small spatial region $Z(\nu)$ in the jet at the scale given by equation (11.31). The volume of this region can be approximated by a cylinder, $V = \pi R^2 Z$, where R is given by equation (11.13). This volume has to be multiplied by the emission coefficient (equation (11.26)) with the magnetic field

(equation (11.18)) inserted. Divided by the surface of an imaginary absorbing sphere at the observer distance D, we get the flux density of the jet as a function of frequency:

$$S_\nu = \epsilon_\nu(\nu_c)\frac{\pi R^2 Z}{4\pi D^2}$$

$$= 1.0\,\mathrm{Jy}\,k\frac{\mathcal{M}}{4}\left(\frac{\gamma_e}{100}\right)^{-1}(\beta_j\gamma_j)^{-3/2}\left(\frac{L_B}{1000L_\odot}\right)^{3/2}. \qquad (11.32)$$

As we can see, the frequency cancels out thus implying a perfectly flat spectrum ($S\nu \propto \nu^0$). This is a fairly general result that applies to essentially all flat-spectrum radio cores in black hole jets. A schematic view of how the flat spectrum arises is also shown in figure 11.9.

The equations have all been normalized by a magnetic luminosity of $L_B \simeq 1000\,L_\odot$. Of course, the jet also has other energetic components (e.g. turbulence or cold protons); however, for a maximally efficient jet they will be of similar order of magnitude (Falcke and Biermann 1995).

The total energy content of the jet will therefore be a few times higher, i.e. of order 10^{37} erg s^{-1}. Assuming the efficiency for jet production to be of order $0.1\dot{M}c^2$ (e.g. Falcke *et al* 1995) this would require an accretion rate of at least $\gtrsim 2 \times 10^{-9}\,M_\odot$ yr^{-1} onto Sgr A*. This is quite in the range of—and sometimes well below—the accretion rates discussed for Sgr A*.

In our derivation we have so far completely ignored relativistic effects on the emitted spectrum. For a continuous relativistic jet where $\beta_j \simeq 1$ and $\gamma_j \gg 1$, relativistic beaming will lead to a modification of the observed spectrum (Lind and Blandford 1985):

$$S'_\nu = \mathcal{D}^2 S_\nu \qquad \text{and} \qquad \nu' = \mathcal{D}\nu \qquad (11.33)$$

where one defines the relativistic Doppler factor as

$$\mathcal{D} = \frac{1}{\gamma_j(1 - \beta_j\cos\theta)}. \qquad (11.34)$$

For moderate inclination angles and moderate jet velocities the effect will be of order unity and we have neglected this for the sake of clarity. From this simple exercise we can conclude that the basic properties of Sgr A*, spectrum, flux, and size, can be naturally reproduced by a jet model.

For a more realistic model, one has to take several additional effects into account. For example, we have here assumed a constant velocity for the jet. This is somewhat inconsistent, since the jet as presented here has a pressure gradient that will naturally lead to some mild acceleration of the jet plasma along the jet axis. This effect together with the relativistic corrections will lead to a slightly inverted radio spectrum and a slightly smaller exponent for the size–frequency relation (Falcke 1996a).
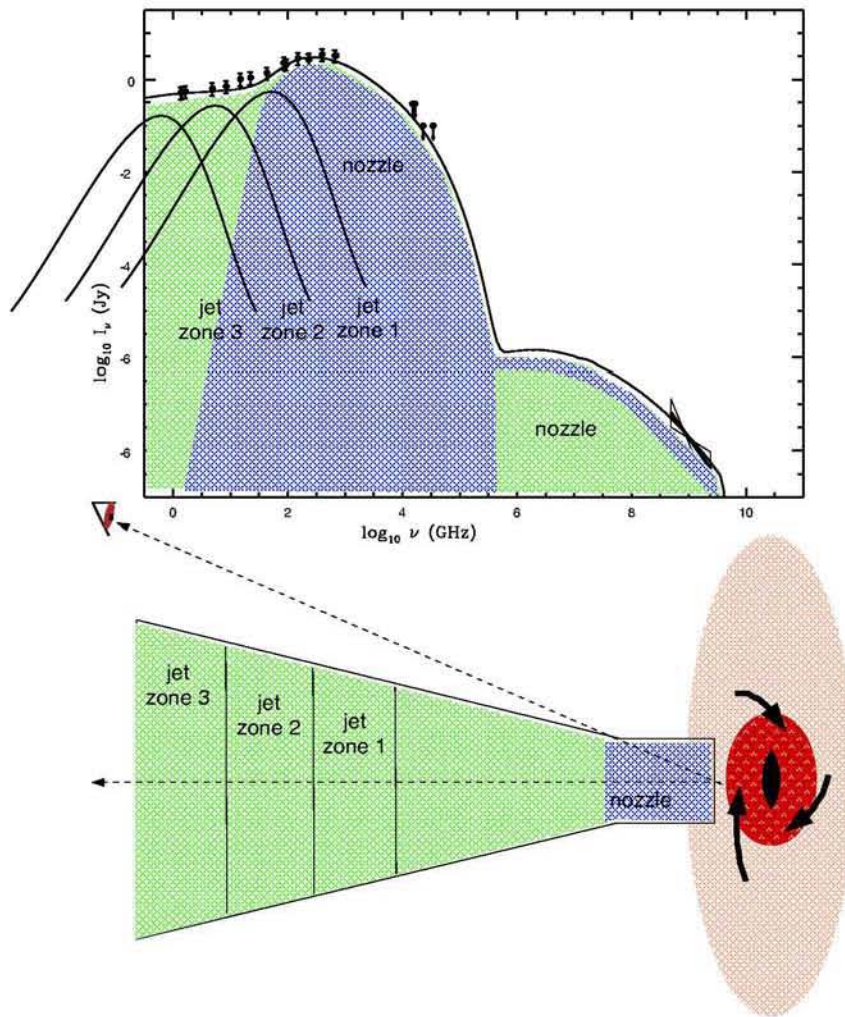
Figure 11.9. This is a schematic diagram of the jet model. The accretion disk/flow is assumed to be radiatively unimportant. The sub-mm bump and the X-ray emission are produced by the jet nozzle region. The flat part of the radio spectrum is the sum of individual, peaked synchrotron spectra from increasingly distant zones in the jet. The peak frequency of each of these spectra shifts to lower frequencies as one moves outwards, while the peak flux density stays essentially at the same level or decreases only slowly. Figure from S Markoff, based on Falcke and Markoff (2000). See also color section.

In addition we have not yet dealt with the sub-mm bump in the spectrum. From equations (11.30) and (11.31) it is clear that this emission has to come from the innermost region of the jet (the 'nozzle'; Falcke 1996b) or the inner edge of the accretion flow (Melia *et al* 2000, Narayan *et al* 1998). A simple descriptive calculation to estimate the parameters of this region, which essentially follows the procedure outlined here, is given in Melia and Falcke (2001, section 5.2).

All these effects including the X-ray spectrum are dealt with in more sophisticated numerical calculations discussed in the next sections.

11.3.2 The X-ray spectrum

After calculating the radio spectrum, we can now make a rough estimate of the expected X-ray spectrum from Sgr A*. For AGN, there are typically four processes discussed to explain the observed X-ray emission in various objects:

(1) synchrotron emission,
(2) Bremsstrahlung,
(3) thermal Comptonization by a hot corona, and
(4) inverse Compton scattering of photons off the relativistic electrons in the jet plasma.

Possibility (1) can be excluded here since the radio synchrotron spectrum cuts off already in the mid-infrared; (2) and (3) have been discussed in chapter 10. Thus, we will here concentrate on the fourth possibility.

Since the only photons we see from Sgr A* outside the X-ray regime are radio photons, here we will consider solely the synchrotron self-Compton (SSC) process, which is absolutely unavoidable. The relativistic electrons that produce synchrotron radiation also have a finite probability to Compton up-scatter the very photons they have produced in the first place. The frequency of the up-scattered photons will be increased by a factor $\simeq \gamma^2$ with respect to the target photons. Inverse Compton is a scattering process where the probability of an interaction of an electron from a population with particle density n_e with a photon of a population with photon density n_γ depends on $n_e \times n_\gamma$. Since in SSC the electrons are also responsible for the target photons, the efficiency of SSC will go as n_e^2. For the case of a jet, where the density increases inwards with R^{-2}, while the volume decreases inwards with R^3, the dominant contribution to the up-scattered spectrum will be at the smallest scale in the system, where n_e is maximal. Following the previous discussion, this will be at a few Schwarzschild radii where the sub-mm bump in the spectrum is produced.

One can show (Rybicki and Lightman 1979, section 7.2) that the luminosity L_{SSC} of the inverse-Compton process is proportional to the luminosity of the synchrotron emission L_{sync}, with the proportionality factor given by the ratio of the energy densities of synchrotron photons:

$$U_{ph} = \frac{L_{sync}}{4\pi R^2 c} \tag{11.35}$$

and magnetic field

$$U_{\rm B} = B^2/8\pi \tag{11.36}$$

such that

$$L_{\rm SSC} = \frac{U_{\rm ph}}{U_{\rm B}} L_{\rm sync}. \tag{11.37}$$

From figure 11.7 we find that the maximum of emission in Sgr A* is about 3 Jy at 10^{12} Hz. Hence the synchrotron luminosity of Sgr A* is

$$L_{\rm sync} = 2.3 \times 10^{35} \frac{\rm erg}{\rm s} \left(\frac{S_\nu}{3\,\rm Jy}\right) \left(\frac{D}{8\,\rm kpc}\right)^2 \left(\frac{\nu_{\rm max}}{10^{12}\,\rm Hz}\right) \tag{11.38}$$

and from equations (11.18) and (11.37) we get, independent of the radius,

$$L_{\rm SSC} = 3.4 \times 10^{33} \frac{\rm erg}{\rm s} (\beta_{\rm j}\gamma_{\rm j}) \left(\frac{S_\nu}{3\,\rm Jy}\right)^2 \left(\frac{D}{8\,\rm kpc}\right)^4 \left(\frac{\nu_{\rm max}}{10^{12}\,\rm Hz}\right)^2 \left(\frac{L_{\rm B}}{1000L_\odot}\right)^{-1}. \tag{11.39}$$

As can be seen the SSC emission is sensitive to the ratio between synchrotron emission and magnetic field. If a different parametrization is used we have a dependence on the equipartition factor k. In general one can also state that the SSC emission should be more variable than the synchrotron emission since it depends with a high power on flux density and peak synchrotron frequency.

The peak of the SSC emission itself will roughly occur at $\gamma^2 \times \nu_{\rm max}$. For $\gamma_{\rm e} \simeq 100$ and $\nu_{\rm max} \simeq 10^{12}$ Hz, the peak will be above 10^{16} eV, hence in the far ultraviolet and soft X-rays.

All of this is quite consistent with the X-ray observations by Baganoff *et al* (2001) who find a quiescent, soft X-ray emission of a few times 10^{33} erg s^{-1} in Sgr A* which can vary rapidly at times.

A schematic view of how the spectrum of Sgr A* from radio to X-rays can be composed from the various parts discussed here is shown in figure 11.9.

11.3.3 Numerical results

Now we have verified that the basic properties of Sgr A* can be explained with a synchrotron+SSC jet model, we can consider a more sophisticated numerical approach. This has been outlined in Falcke and Markoff (2000), Markoff *et al* (2001), and Yuan *et al* (2002).

We start with the basic jet emission model (Falcke and Biermann 1999, Falcke and Markoff 2000), consisting of a conical jet with pressure gradient, nozzle and relativistic effects. The parameters in the nozzle for the quiescent state are determined from the underlying accretion disk, assumed to be an ADAF, as described in Yuan *et al* (2002). All quantities further out in the jet are solved by using conservation of mass and energy, and the Euler equation for the accelerating velocity field. The results are shown in figures 11.10 and 11.11 and show that the model is able to reproduce the observed spectrum and size in detail.
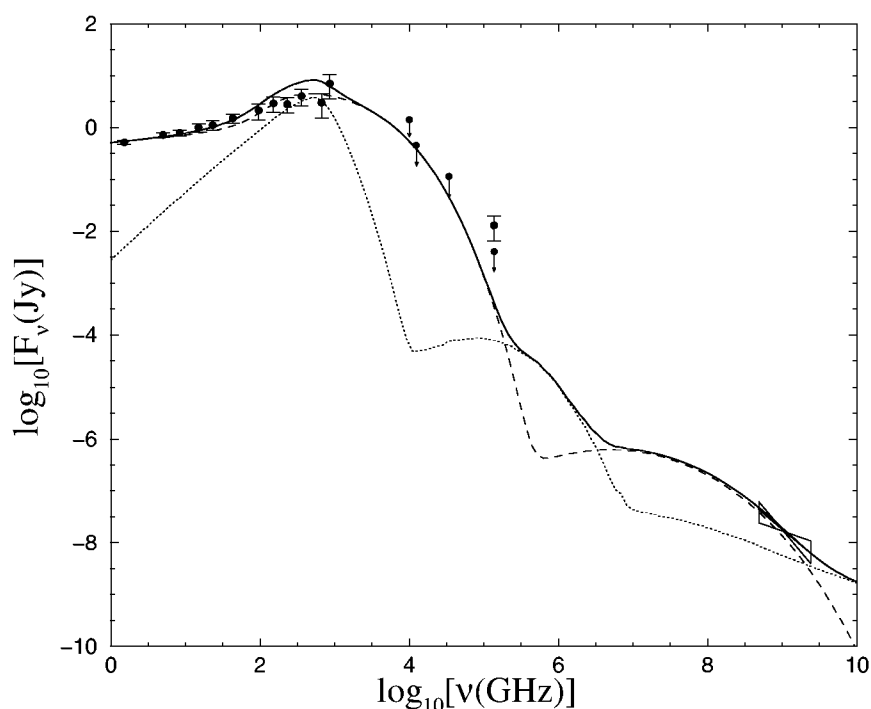
Figure 11.10. The jet–disk spectral model for Sgr A*. The dotted line is for the ADAF (optically thin, advection dominated accretion flow) contribution, the dashed line is for the jet emission, and the full line shows their sum. For the most part, the emission is dominated by the jet spectrum. The sub-mm bump is produced by the jet nozzle with a possible contribution from the accretion flow. The X-ray emission is largely SSC emission from the nozzle with a slight contribution from the more extended thermal X-ray emission from the accretion flow. We have here assumed an accretion rate of 10^{-6} $M_\odot$ yr^{-1}, where only 0.1% of the power goes into the jet. For a 10% efficiency the required accretion rate is about 10^{-8} $M_\odot$ yr^{-1} and the disk contribution would be negligible. The model is discussed in more detail in Yuan *et al* (2002).

11.3.4 The circular polarization

Finally, to understand the radio properties we also have to consider the surprising results of the polarization observations, where a relatively large circular polarization (CP) was found. The following intuitive explanation is essentially a discussion of CP based on the paper by Beckert and Falcke (2002), where more details can be found. The main point is that linear polarization is naturally obtained in synchrotron radiation (up to 70% for homogeneous magnetic fields), while CP is strongly suppressed in synchrotron plasmas with $\gamma_e \gg 1$. The reason
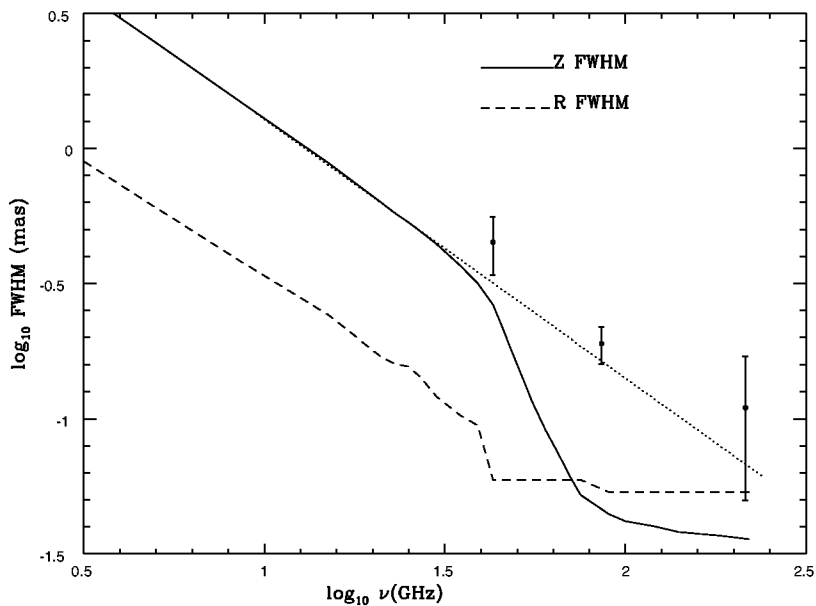
Figure 11.11. Projected size of the major and minor axes of the jet in Sgr A* as a function of frequency. The filled dots mark the size as measured by Lo *et al* (1998; 43 GHz) and Krichbaum *et al* (1998; 86 and 215 GHz). The lines represent the predictions of the jet model. At frequencies above 30 GHz one obtains a two-component structure with an increasingly stronger core (nozzle, full dashed line) and a fainter jet component (dotted line).

for this is that the narrow beaming cone of the relativistic electrons allows one to see only a small arc along the gyration around the magnetic field. However, CP can be obtained through radiation transfer, particularly through the fact that a magnetic plasma will be naturally birefringent.

For simplicity let us now separate Faraday rotation from conversion and picture only purely linearly or circularly polarized waves in a homogeneous magnetic field.

The two orthogonal normal modes for propagation perpendicular to the magnetic field are linearly polarized and a purely circularly polarized wave is split into the two normal modes with a relative phase-shift as shown in figure 11.12 (left). Without a phase-shift the wave will be purely linearly polarized. If, for example, a locally homogeneous magnetic field vertically pervades the box in figure 11.12 (left) along the z-direction, electrons or positrons will be free to move along the field lines and resonate with the vertical mode but hardly resonate with the horizontal mode along the x-direction. This yields the birefringence discussed
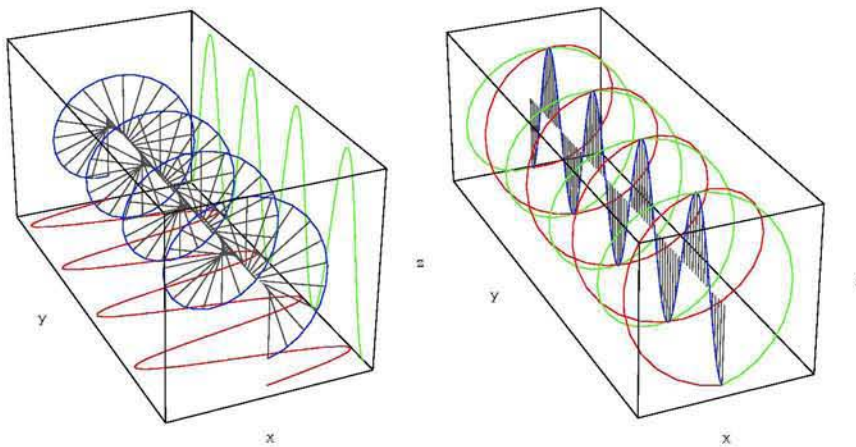
Figure 11.12. Left: A circularly polarized wave can be composed of two orthogonal linearly polarized modes shifted in phase. A phase shift would be produced by a plasma in a magnetic field perpendicular to the propagation direction of the waves (here along the z-direction). Without a phase-shift the sum of the two modes would be a purely linearly polarized wave. A video showing the effect of how phase-shifts in a region will turn such a linearly polarized wave into a circularly polarized wave (conversion) can be found on http://bookmarkphysics.iop.org/bookpge.htm?&book=1153p. Right: A linearly polarized wave can be composed of two orthogonal circularly polarized modes shifted in phase. A phase shift would be produced by a plasma in a magnetic field along the propagation direction of the waves (here along the y-direction). A video showing the effect of additional phase-shifts on the linear polarization, leading to Faraday rotation, can be found at http://bookmarkphysics.iop.org/bookpge.htm?&book=1153p.

earlier. The resonating electrons or positrons will themselves act as antennas and emit a somewhat delayed wave that interferes with the incoming vertical mode, leading to a slight phase-shift between vertical and horizontal modes. The effect of this shift is shown in the animation[4], where the resulting wave is circularly polarized and switches from linear to circular polarization as a function of the shift.

Conversion also acts on initially only linearly polarized radiation. The extent of this conversion will depend on the misalignment between the incoming wave and the magnetic field direction since, obviously, a phase-shift between two orthogonal modes will have little effect if one mode is very small or non-existent. Moreover, a random distribution of magnetic field lines on the plane of the sky will reduce circular polarization from conversion in exactly the same way as linear polarization would be reduced.

Analogous to the picture for conversion, one can view a linearly polarized

---

[4] http://bookmarkphysics.iop.org/bookpge.htm?&book=1153p

wave as composed of two circularly polarized normal modes when propagating along the magnetic field. This is sketched in figure 11.12 (right), where we will assume a longitudinal magnetic field, i.e. a field along the y-direction. The circular modes will resonate with either electrons or positrons gyrating around the magnetic fields. The latter will again emit a circularly polarized wave, producing a phase-shift when interfering with the incoming wave. The effect of the phase-shift in the circular modes is shown in an animation of figure 11.12 (right), where one can see that the resulting linearly polarized wave is simply (Faraday) rotated (see http://bookmarkphysics.iop.org/bookpge.htm?&book=1153p).

An important conclusion to remember, therefore, is that conversion is mainly produced by magnetic field components perpendicular to the line-of-sight or photon direction, while Faraday rotation is produced by magnetic field components along the line-of-sight. Moreover, one can also see that conversion is insensitive to the electron/positron ratio while Faraday rotation is not. In figure 11.12 (left) an electron and an positron are both free to move along the z-axis. While they will respond in opposite directions to the incoming wave, their respective emitted waves will also have opposite signs because of opposite charges and hence be identical. In the case of Faraday rotation, the incoming left- or right-handed circularly polarized wave will only resonate with the particle that also has the correct handedness in its gyration—either electron or positron depending on the magnetic field polarity. A pure pair plasma would therefore produce exactly the same phase-shift in left- and right-handed modes and not produce any net Faraday rotation. In the case of excess charge, the direction of Faraday rotation depends on the sign of the excess charge (presumably electrons) and the polarity of the magnetic field. This will indirectly also affect the sign of the circular polarization, if Faraday rotation is the ultimate cause of the misalignment between the plane of polarization and the magnetic field direction.

One can include these effects on the polarization into a radiation transfer code and try to reproduce the Sgr A* spectrum and polarization with a jet/outflow model (Beckert and Falcke 2002). The results are shown in figure 11.13 and nicely reproduce the observed spectrum. Two major conclusions can be drawn from this approach:

(i)   Since conversion is most effective for low-energy electrons one can conclude that a larger number of these low-energy electrons ($1 < \gamma_e \ll 100$) are present in Sgr A*. In fact, in the specific modeling mentioned here, one finds that up to a factor of 100 more low-energy electrons could be present than the 'hot' electrons with $\gamma_e \simeq 100$ invoked earlier to explain the spectrum. Hence, a large fraction of Sgr A*'s plasma could reside in a rather inconspicuous, 'cold' and non-radiating (hidden) plasma.

(ii) Since the conversion requires an asymmetric magnetic field polarity, an outflow model with a helical magnetic field is strongly favored. The stability of the sign of the CP also suggests that the polarity (the North pole of the black hole/jet) has remained stable over some 20 years—a long timescale
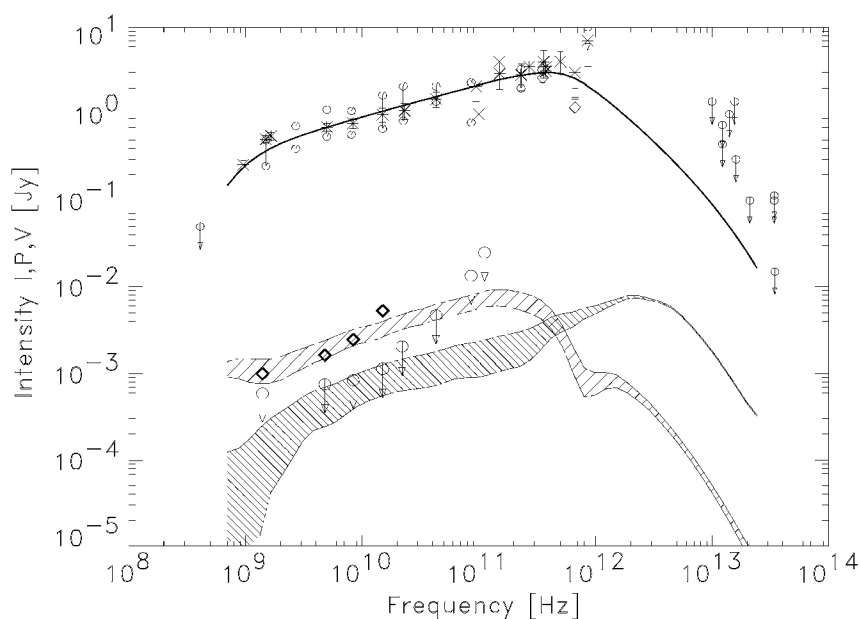
Figure 11.13. Outflow model for the radio spectrum of Sgr A* with polarization. The result of model calculations for total flux I (full line), linearly polarized P (dense shaded area), and circularly polarized flux V (sparsely shaded area) are shown for a distance of 8 kpc. Diamonds show the observed circular polarization and circles the observed linear polarization; the rest are observed total flux density values. The numerical calculations are based on the model described in Beckert and Falcke (2002). The shaded areas mark the expected variability due to turbulence. The global magnetic field structure is a spiral with $B_\phi/B_z = 1$.

compared to the accretion time in optically thin accretion flows. It is possible that this stable polarity reflects the stable polarity of the Galactic magnetic field pervading the central parsec of the Galaxy which is accreted onto the black hole.

The main uncertainty in the modeling of the polarization at present reflects the uncertainty in what suppresses the linear polarization in Sgr A*. Here, we have simply assumed that the de-polarization of the linear polarization is due to intrinsic Faraday de-polarization in the radio source itself. However, this could similarly be done in a 'foreground' screen, which would most likely be associated with the accretion flow. If that is the case, some of the 'hidden matter' mentioned earlier would be in the actual accretion flow and not in the outflow/jet. This can be used to constrain the accretion rate. Estimates by Agol (2000) and Quataert and Gruzinov (2000) then yield a limit on the accretion flow of $\dot{M} \lesssim 10^{-8}$ to

10^{-9} $M_\odot$ yr^{-1}, given the rotation measures inferred from the linear polarization observations at a λ of 1 mm.

11.3.5   Comparison with other supermassive black holes

An important input factor to the model discussed so far is the power of the jet (here we mainly considered the magnetic power L_B). A nice feature of the model is that it can be scaled over many orders of magnitude by just changing the power input of the jet. This is presumably done by a parallel change in the accretion disk power. Doing so would change the power of the radio core but would not change its spectral shape—only the turnover frequency might change. To a radio observer the jet would always appear as a flat-spectrum core. This may be the reason why radio jets are expected for almost every type of active black hole: from supermassive to stellar, from powerful to faint.

Indeed, compact, flat-spectrum radio cores have been found in sources like quasars, Seyfert galaxies, low-luminosity AGNs and LINERs, as well as in X-ray binaries, confirming that the physics we have discussed for Sgr A* is fairly universal. The exact nature of the cores and the emission of these other engines is not the main focus of the book and further discussion of this point can be found in Falcke (2001). However, the general point one can make is that as the accretion power and the disk luminosity decreases, one expects to see fainter radio cores. If there is a range in accretion rates throughout the universe, we also expect a range of core luminosities. This is demonstrated in figure 11.14, where Sgr A*-like radio cores of various different types are shown for a range of luminosities. In such a plot, Sgr A* would come in at the bottom left of the distribution for an accretion rate of $\simeq 10^{-8}$ $M_\odot$ yr^{-1} (as indicated by the upper black dot with a horizontal error bar; the other point indicates the estimated position for M31*—the core in the Andromeda galaxy). However, since the accretion disk in Sgr A* is so faint and the accretion rate so uncertain, we cannot actually derive an accretion disk luminosity or accretion power and this should only be considered as a general guideline. The bottom line is, however, that Sgr A* with its radio properties is not alone in the universe but is at the bottom end of the activity scale seen from supermassive black holes.

11.4   Imaging the event horizon—an outlook

One can easily see from the previous sections that the ever growing interest in Sgr A* has already yielded a number of tantalizing results, the most important being that Sgr A* is the best supermassive black hole candidate we know. VLBI observations are already approaching scales which are not far from the actual scale of the black hole and the presence of the sub-mm bump indicates that even more compact emission is present at yet smaller scales—possibly as close in as the event horizon of the black hole. It is therefore worth exploring whether we have, in principle, a chance to actually approach this scale with imaging techniques
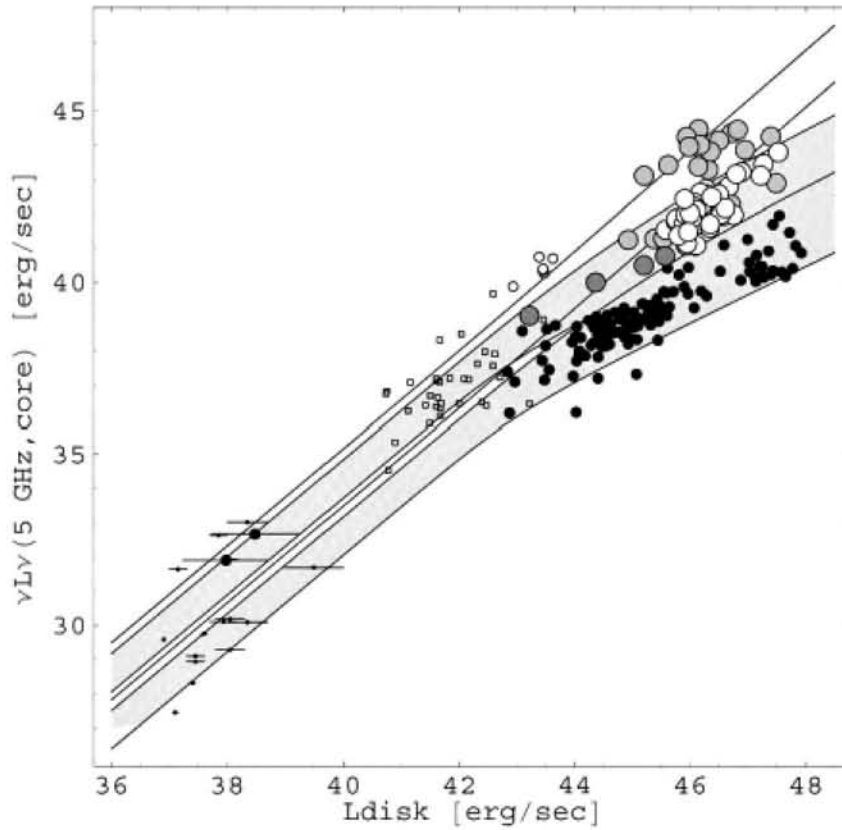
Figure 11.14. Correlation between thermal from the accretion disk (with the exception of X-ray binaries this is basically normalized to the narrow Hα emission) and the monochromatic luminosity of black hole radio cores: open circles, radio-loud quasars; small open circles, FR I radio galaxies; open gray circles, blazars and radio-intermediate quasars (dark grey); black dots, radio-quiet quasars and Seyferts; small dots, X-ray binaries; small boxes, detected sources from the '48 LINERs' sample (Nagar *et al* 2000). The latter apparently confirm the basic prediction of Falcke and Biermann (1996) and almost close the gap between very low (on an absolute scale) accretion rate objects and high accretion rate objects. The shaded bands represent the radio-loud and radio-quiet jet models as a function of accretion as shown in Falcke and Biermann (1996).

and to ask what we would expect to see. This naturally will have to be done at the highest radio frequencies where the resolution is the highest and the scatter-broadening of Sgr A* is the lowest.

At sub-mm wavelengths, the various models indeed predict that the synchrotron emission of Sgr A* is not self-absorbed, allowing a view into the

region near the event horizon. The size of this event horizon is $\frac{1}{2}(1+\sqrt{1-a_*^2})R_S$, where $R_S \equiv 2GM/c^2$, M is the mass of the black hole, G is Newton's constant, c the speed of light, $a_* \equiv Jc/(GM^2)$ is the dimensionless spin of the black hole in the range 0–1, and J is the angular momentum of the black hole.

The appearance of the emitting region around a black hole was determined by Falcke *et al* (2000)—from which we take the following discussion—under the condition that it is optically thin. For Sgr A* this might be the case for the sub-mm bump (Falcke *et al* 1998) indicated by the turnover in the spectrum, and can always be achieved by going to a suitably high frequency. For the qualitative discussion the emissivity was assumed to be frequency independent and to be either spatially uniform or to scale as r^{-2}. These two cases cover a large range of conditions expected under several reasonable scenarios, be it a quasi-spherical infall, a rotating thick disk, or the base of an outflow. The calculations took into account all the well-known relativistic effects, e.g. frame dragging, gravitational redshift, light bending, and Doppler boosting.

For a planar emitting source behind a black hole, a closed curve on the sky plane divides a region where geodesics intersect the horizon from a region whose geodesics miss the horizon (Bardeen 1973)[5]. This curve, which is referred to as the 'apparent boundary' of the black hole, is a circle of radius $\frac{1}{2}\sqrt{27}R_S$ in the Schwarzschild case ($a_* = 0$), but has a more flattened shape of similar size for a Kerr black hole, slightly dependent on inclination. The size of the apparent boundary is much larger than the event horizon due to strong bending of light by the black hole. When the emission occurs in an optically thin region *surrounding* the black hole, the case of interest here, the apparent boundary has the same exact shape since the properties of the geodesics are independent of where the sources are located. However, photons on geodesics located within the apparent boundary that can still escape to the observer experience strong gravitational redshift and a shorter total path length, leading to a smaller integrated emissivity, while photons just outside the apparent boundary can orbit the black hole near the circular photon radius several times, adding to the observed intensity (Jaroszynski and Kurpiewski 1997). This produces a marked deficit of the observed intensity inside the apparent boundary—the 'shadow' of the black hole.

We here consider a compact, optically thin emitting region surrounding a black hole with spin parameter $a_* = 0$ (i.e. a Schwarzschild black hole) and a maximally spinning Kerr hole with $a_* = 0.998$. In the set of simulations shown in figure 11.15, the viewing angle i was taken to be 45° with respect to the spin axis (when it is present) with two distributions of the gas velocity v. The first has the plasma in free fall, i.e. $v^r = -\sqrt{2r(a^2 + r^2)}\Delta/A$ and $\Omega = 2ar/A$, where v^r is the Boyer–Lindquist radial velocity, Ω is the orbital frequency, $\Delta \equiv r^2 - 2r + a^2$, and $A \equiv (r^2 + a^2)^2 - a^2\Delta\sin^2\theta$. (We have set $G = M = c = 1$ in this paragraph.) The second has the plasma orbiting in rigidly rotating shells with the equatorial Keplerian frequency $\Omega = 1/(r^{3/2} + a)$ for $r > r_{ms}$ (marginally

---

[5] For the following discussion see also chapter 5.
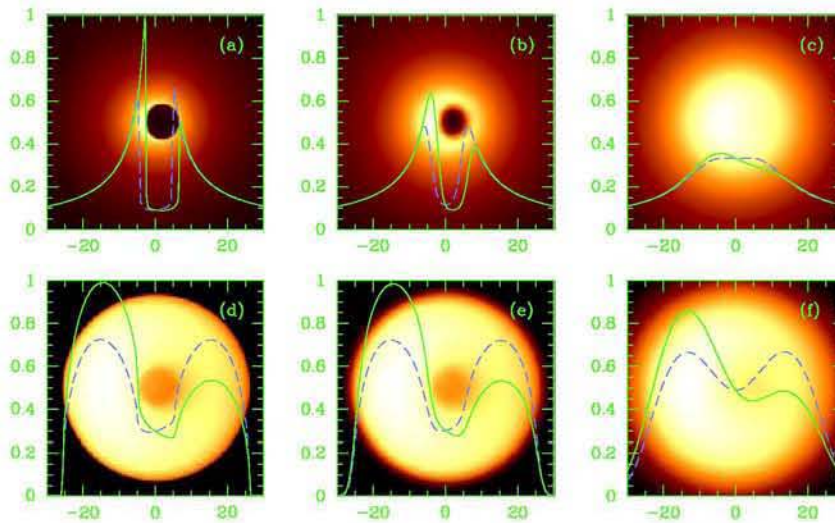
Figure 11.15. An image of an optically thin emission region surrounding a black hole with the characteristics of Sgr A* at the Galactic Center. The black hole is here either maximally rotating ($a_* = 0.998$, panels (a)–(c)) or non-rotating ($a_* = 0$, panels (d)–(f)). The emitting gas is assumed to be in free fall with an emissivity $\propto r^{-2}$ (top) or on Keplerian shells (bottom) with a uniform emissivity (viewing angle $i = 45°$). Panels (a) and (d) show the GR ray-tracing calculations, panels (b) and (e) are the images seen by an idealized VLBI array at 0.6 mm wavelength taking interstellar scattering into account, and panels (c) and (f) are those for a wavelength of 1.3 mm. The intensity variations along the x-axis (full green curve) and the y-axis (dashed purple/blue curve) are overlaid. The vertical axes show the intensity of the curves in arbitrary units and the horizontal axes show the distance from the black hole in units of $GM_\bullet/c^2$ ($1/2R_S$). See also color section and video at http://bookmarkphysics.iop.org/bookpge.htm?&book=1153p.

stable radius) with $v^r = 0$, and infalling with constant angular momentum inside $r < r_{ms}$ (Cunningham 1975), with $v^\theta = 0$ for all r.

In order to display concrete examples of how realistic the proposed measurements of these effects with VLBI will be, the expected images were simulated for the massive black hole candidate Sgr A* at the Galactic Center.

The results of the two different models with and without scattering at two different observing wavelengths are shown in figure 11.15. The two distinct features that are evident in the top panel for a rotating black hole are: (1) the clear depression in I_ν—the shadow—produced near the black hole, which in this particular example represents a modulation of up to 90% in intensity from peak to trough; and (2) the size of the shadow, which here is 4.6 R_S in diameter. This represents a projected size of 34 μarcseconds. Such a resolution has already been surpassed in some λ2 mm-VLBI experiments of other radio cores (Krichbaum *et*

al 2002). The shadow is a generic feature of various other models one can look at, including those with outflows, cylindrical emissivity, and various inclinations or spins.

This black hole shadow is also visible in the second illustrated case for a non-rotating black hole with a modulation in I_ν in the range of 50–75% from peak to trough, and with a diameter of roughly 5.2 R_S. In this case, the emission is asymmetric due to the strong Doppler shifts associated with the emission by a rapidly moving plasma along the line-of-sight (with velocity v_ϕ).

The important conclusion is that the diameter of the shadow—in marked contrast to the event horizon—is fairly independent of the black hole spin and is always of order 5 R_S. The presence of a rotating hole viewed edge-on will lead to a shifting of the apparent boundary (by as much as 1.25 R_S, or 9 μarcseconds) with respect to the center of mass, or the centroid of the outer emission region. Another possible signature of general relativistic effects may come from the polarization properties of the sub-mm-wave emission region. This has been calculated by Bromley *et al* (2001).

The importance of the proposed imaging of Sgr A* at sub-mm wavelengths with VLBI cannot be overemphasized. The sub-mm bump in the spectrum of Sgr A* strongly suggests the presence of a compact component whose proximity to the event horizon is predicted to result in a shadow of measurable dimensions in the intensity map. Such a feature seems unique and Sgr A* seems to have all the right parameters to make it observable. The observation of this shadow would confirm the widely held belief that most of the dark mass concentration in the nuclei of galaxies such as ours is contained within a black hole, and it would be the first direct evidence for the existence of an event horizon largely independent of any modeling. *It would finally turn the theoretical concept of an event horizon discussed at the beginning of the book into an observable reality.*

A non-detection with sufficiently developed techniques, however, might pose a major problem for the standard black hole paradigm. Because of this fundamental importance, the experiment proposed here should be a major motivation for intensifying the current development of sub-mm astronomy in general and mm and sub-mm VLBI in particular.

This result also shows the outstanding position Sgr A* has among known radio cores and supermassive black hole candidates. For other supermassive black holes, with the exception perhaps of the very massive black hole in M87, the shadow will be much smaller than in Sgr A* because of the much larger distances.

References

Agol E 2000 *Astrophys. J.* **538** L121
Aitken D K, Greaves J, Chrysostomou A, Jenness T, Holland W, Hough J H, Pierce-Price D and Richer J 2000 *Astrophys. J.* **534** L173
Backer D C and Sramek R A 1999 *Astrophys. J.* **524** 805
Baganoff F K *et al* 2001 *Nature* **413** 45

Balick B and Brown R L 1974 *Astrophys. J.* **194** 265

Bardeen J M 1973 *Black Holes* ed C DeWitt and B S DeWitt (New York: Gordon and Breach) p 215

Beckert T and Falcke H 2002 *Astron. Astrophys.* **388** 1106

Bietenholz M F, Bartel N and Rupen M P 2000 *Astrophys. J.* **532** 895

Bower G C and Backer D C 1998 *Astrophys. J.* **496** L97

Bower G C, Backer D C, Zhao J H, Goss M and Falcke H 1999a *Astrophys. J.* **521** 582

Bower G C, Falcke H and Backer D C 1999b *Astrophys. J.* **523** L29

Bower G C, Falcke H, Sault R J and Backer D C 2002a *Astrophys. J.* **571** 843

Bower G C, Wright M C H, Falcke H and Backer D C 2002b *Astrophys. J.* submitted

Bridle A H and Perley R A 1984 *Annu. Rev. Astron. Astrophys.* **22** 319

Bromley B C, Melia F and Liu S 2001 *Astrophys. J.* **555** L83

Brown R L and Lo K Y 1982 *Astrophys. J.* **253** 108

Brunthaler A, Bower G C, Falcke H and Mellon R R 2001 *Astrophys. J.* **560** L123

Cunningham C T 1975 *Astrophys. J.* **202** 788

Davies R D, Walsh D and Booth R S 1976 *Mon. Not. R. Astron. Soc.* **177** 319

Doeleman S S *et al* 2001 *Astrophys. J.* **121** 2610

Falcke H 1996a *Astrophys. J.* **464** L67

——1996b *Unsolved Problems of the Milky Way (IAU Symp. 169)* ed L Blitz and P Teuben (Dordrecht: Kluwer) p 169

——1999 *The Central Parsecs of the Galaxy (ASP Conf. Ser. 186)* ed H Falcke, A Cotera, W Duschl, F Melia and M J Rieke (San Francisco: Astronomical Society of the Pacific) p 113

——2001 *Rev. Mod. Astron.* **14** 15

Falcke H and Biermann P L 1994 *Mass Transfer Induced Activity in Galaxies* ed I Shlosman (Cambridge: Cambridge University Press) p 44

——1995 *Astron. Astrophys.* **293** 665

——1996 *Astron. Astrophys.* **308** 321

——1999 *Astron. Astrophys.* **342** 49

Falcke H, Goss W M, Matsuo H, Teuben P, Zhao J H and Zylka R 1998 *Astrophys. J.* **499** 731

Falcke H, Malkan M A and Biermann P L 1995 *Astron. Astrophys.* **298** 375

Falcke H and Markoff S 2000 *Astron. Astrophys.* **362** 113

Falcke H, Melia F and Agol E 2000 *Astrophys. J.* **528** L13

Falcke H, Nagar N M, Wilson A S and Ulvestad J S 2000 *Astrophys. J.* **542** 197

Ferrari A 1998 *Annu. Rev. Astron. Astrophys.* **36** 539

Jaroszynski M and Kurpiewski A 1997 *Astron. Astrophys.* **326** 419

Königl A 1980 *Phys. Fluids* **23** 1083

Krichbaum T *et al* 1993 *Astron. Astrophys.* **274** L37

——1998 *Astron. Astrophys.* **335** L106

——2002 *Proc. 6th Eur. VLBI Network Symp.* ed E Ros, R Porcas, A Lobanov and J Zensus (Bonn: Max-Planck-Insitut für Radioastronomie) p 125

Krichbaum T P, Witzel A and Zensus J A 1999 *The Central Parsecs of the Galaxy (ASP Conf. Ser. 186)* ed H Falcke, A Cotera, W Duschl, F Melia and M J Rieke (San Francisco: Astronomical Society of the Pacific) p 89

Lind K R and Blandford R D 1985 *Astrophys. J.* **295** 358

Liu S and Melia F 2002 *Astrophys. J.* **573** L23

Lo K Y, Shen Z, Zhao J H and Ho P T P 1998 *Astrophys. J.* **508** L61

——1999 *The Central Parsecs of the Galaxy (ASP Conf. Ser. 186)* ed H Falcke, A Cotera, W Duschl, F Melia and M J Rieke (San Francisco: Astronomical Society of the Pacific) p 72

Lynden-Bell D and Rees M J 1971 *Mon. Not. R. Astron. Soc.* **152** 461

Ma C, Arias E F, Eubanks T M, Fey A L, Gontier A-M, Jacobs C S, Sovers O J, Archinal B A and Charlot P 1998 *Astrophys. J.* **116** 516

Marcaide J M, Alberdi A, Lara L, Pérez-Torres M A and Diamond P J 1999 *Astron. Astrophys.* **343** 801

Markoff S, Falcke H, Yuan F and Biermann P L 2001 *Astron. Astrophys.* **379** L13

Melia F and Falcke H 2001 *Annu. Rev. Astron. Astrophys.* **39** 309

Melia F, Liu S and Coker R 2000 *Astrophys. J.* **545** L117

Menten K M, Reid M J, Eckart A and Genzel R 1997 *Astrophys. J.* **475** L111

Nagar N M, Falcke H, Wilson A S and Ho L C 2000 *Astrophys. J.* **542** 186

Narayan R, Mahadevan R, Grindlay J E, Popham R G and Gammie C 1998 *Astrophys. J.* **492** 554

Pacholczyk A G 1970 *Radio Astrophysics* (San Francisco: Freeman)

Palmer P and Goss W M 1996 *Galactic Center News.* **2** 3

Pedlar A, Anantharamaiah K R, Ekers R D, Goss W M, van Gorkom J H, Schwarz U J and Zhao J 1989 *Astrophys. J.* **342** 769

Quataert E and Gruzinov A 2000 *Astrophys. J.* **545** 842

Reid M J, Readhead A C S, Vermeulen R C and Treuhaft R N 1999 *Astrophys. J.* **524** 816

Reuter H P and Lesch H 1996 *Astron. Astrophys.* **310** L5

Rogers A E E *et al* 1994 *Astrophys. J.* **434** L59

Rybicki G B and Lightman A P 1979 *Radiative Processes in Astrophysics* (New York: Wiley) p 393

Scheuer P 1968 *Nature* **218** 920

Serabyn E, Carlstrom J, Lay O, Lis D C, Hunter T R and Lacy J H 1997 *Astrophys. J.* **490** L77

Tsuboi M, Miyazaki A and Tsutsumi T 1999 *The Central Parsecs of the Galaxy (ASP Conf. Ser. 186)* ed H Falcke, A Cotera, W Duschl, F Melia and M J Rieke (San Francisco: Astronomical Society of the Pacific) p 105

van Langevelde H J, Frail D A, Cordes J M and Diamond P J 1992 *Astrophys. J.* **396** 686

Wright M C H and Backer D C 1993 *Astrophys. J.* **417** 560

Wrobel J M and Heeschen D S 1984 *Astrophys. J.* **287** 41

Yuan F, Markoff S and Falcke H 2002 *Astron. Astrophys.* **383** 854

Yusef-Zadeh F, Choate D and Cotton W 1999 *Astrophys. J.* **518** L33

Yusef-Zadeh F, Cotton W, Wardle M, Melia F and Roberts D A 1994 *Astrophys. J.* **434** L63

Zensus J A 1997 *Annu. Rev. Astron. Astrophys.* **35** 607

Zhao J, Bower G C and Goss W M 2001 *Astrophys. J.* **547** L29

Zhao J-H, Young K H, McGary R S, Ho P T P, Tsutsumi T, Goss W M and Bower G C 2001 (American Astronomical Society Meeting) p 199

Zylka R, Mezger P G, Ward-Thompson D, Duschl W J and Lesch H 1995 *Astron. Astrophys.* **297** 83

Appendix A

List of authors

Dr Tal Alexander
Faculty of Physics
The Weizmann Institute of Science
Rehovot 76100
Israel
Email: tal.alexander@weizmann.ac.il

Dr Robert F Coker
Los Alamos National Laboratory
MS T-087, LANL
Los Alamos, NM 87545
USA
Email: robc@lanl.gov

Prof Andreas Eckart
I. Physikalisches Institut
Universität zu Köln
Zülpicher Straße 77
50937 Köln
Germany
Email: eckart@ph1.uni-koeln.de

343

Prof Heino Falcke
Max-Planck-Institut für Radioastronomie
Auf dem Hügel 69
53121 Bonn
Germany
Email: hfalcke@mpifr-bonn.mpg.de

Prof Valeri Frolov
Department of Physics
University of Alberta
Edmonton AB
Canada T6G 2J1
Email: frolov@phys.ualberta.ca

PD Dr Domenico Giulini
Department of Physics
University of Freiburg
Hermann-Herder-Strasse 3
79104 Freiburg i.Br.
Germany
Email: giulini@physik.uni-freiburg.de

Prof Friedrich W Hehl
Inst. Theor. Physics
University of Cologne
50923 Köln
Germany
Email: hehl@thp.uni-koeln.de

Dipl Phys Christian Heinicke
Inst. Theor. Physics
University of Cologne
50923 Köln
Germany
Email: chh@thp.uni-koeln.de

HD Dr Susanne Hüttemeister
Astronomisches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44780 Bochum
Germany
Email: huette@astro.ruhr-uni-bochum.de

Prof Claus Kiefer
Inst. Theor. Physics
University of Cologne
50923 Köln
Germany
Email: kiefer@thp.uni-koeln.de

Prof Mark R Morris
Department of Physics and Astronomy
University of California, Los Angeles
Los Angeles, CA 90095-1562
USA
Email: morris@astro.ucla.edu
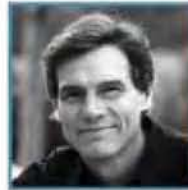
Prof Gernot Neugebauer
Theoretisch-Physikalisches Institut
Friedrich-Schiller-Universität Jena
Max-Wien-Platz 1
07743 Jena
Germany
Email: G.Neugebauer@tpi.uni-jena.de

Appendix B

Units and constants

In table B.1 we list a few constants and units that are commonly used in astrophysics and astronomy. Since astronomers often still use cgs units while physicists use SI units, we use both systems here.

In table B.2 we list a few typical astronomical observing bands with corresponding frequency v, wavelength λ, and energy range E. The boundaries between the various bands are not very stringent.

Astrophysical literature quoted here can be easily found through the NASA Astrophysical Data System (ADS) at http://adswww.harvard.edu. Preprints can be found at http://arxiv.org.

Table B.1. Astrophysical and physical constants and units.

Speed of light:	c	=	2.998×10^8 m s^{-1}	=	2.998×10^{10} cm s^{-1}
Planck's constant:	h	=	6.626×10^{-34} J s	=	6.626×10^{-27} cm^2 g s^{-1}
Gravitational constant:	G	=	6.67×10^{-11} m^3 s^{-2} kg^{-1}	=	6.67×10^{-8} cm^3 g^{-1} s^{-2}
Boltzmann's constant:	k_B	=	1.38×10^{-23} J K^{-1}	=	1.38×10^{-16} cm^2 g s^{-2} K^{-1}
Electron mass:	m_e	=	9.11×10^{-31} kg	=	9.11×10^{-28} g
Proton mass:	m_p	=	1.67×10^{-27} kg	=	1.67×10^{-24} g
Electron charge:	e	=	1.60×10^{-19} A s	=	4.80×10^{-10} cm$^{3/2}$ g$^{1/2}$ s^{-1}
Stefan–Boltzmann constant:	σ	=	5.67×10^{-8} W m^{-2} K^{-4}	=	5.67×10^{-5} g s^{-3} K^{-4}
Thomson cross section:	σ_{th}	=	6.65×10^{29} m^2	=	6.65×10^{-25} cm^2
Solar mass:	$M_\odot$	=	1.99×10^{30} kg	=	1.99×10^{33} g
Solar radius:	$R_\odot$	=	0.696×10^6 km	=	6.96×10^{10} cm
Solar luminosity:	$L_\odot$	=	3.85×10^{26} W	=	3.85×10^{33} erg s^{-1}
Jansky (energy flux density):	1 Jy	=	10^{-26} W s^{-1} cm^{-2} Hz^{-1}	=	10^{-23} erg s^{-1} cm^{-2} Hz^{-1}
Energy:	1 J (= joule)				10^7 erg
Magnetic field B:	1 T (= tesla)				10^4 G (= gauss)
astronomical unit:	1 AU	=	150×10^6 km	=	1.50×10^{13} cm
lightyear:	1 ly	=	0.946×10^{13} km	=	0.946×10^{18} cm
parsec:	1 pc	=	3.09×10^{13} km	=	3.09×10^{18} cm
kiloparsec:	1 kpc	=	3.09×10^{16} km	=	3.09×10^{21} m
arcdegree:	$1°$	=	$2\pi/360$	=	1.75×10^{-2}
arcminute:	$1'$	=	$1°/60$	=	2.91×10^{-4}
arcsecond:	$1''$	=	$1'/60$	=	4.85×10^{-6}
milliarcsecond:	1 mas	=	$1''/1000$	=	4.85×10^{-9}

As an example: the diameter of the full moon on the sky is $31'$, i.e. half a degree.

Note: for parameters of the Galaxy see p 41.

Galactic Center distance: $D_{GC} = 8$ kpc $= 26\,100$ ly $= 2.5 \times 10^{17}$ km.

Mass of Galactic Black Hole: $M_\bullet \simeq 3 \times 10^6\ M_\odot$.

Schwarzschild radius of Galactic Black Hole: $R_S = 2GM_\bullet/c^2 = 8.9 \times 10^6$ km. (Many papers also define a 'gravitational radius', which in general relativity papers is usually synonymous to Schwarzschild radius, while in some astrophysical papers it is defined as $R_g = GM_\bullet/c^2$ (i.e. 4.4×10^9 m for the Galactic Black Hole).)

Table B.2. Astronomical observing bands.

Band	ν	λ	$E = h\nu$
Radio (m wave)	10 MHz–1 GHz	30 m–30 cm	4×10^{-8} eV–4×10^{-6} eV
Radio (cm wave)	1 GHz–100 GHz	30 cm–3 mm	4×10^{-6} eV–4×10^{-4} eV
mm and sub-mm wave	100 GHz–1 THz	3 mm–0.3 mm	4×10^{-4} eV–4×10^{-3} eV
Far-infrared (FIR)	1 THz–10 THz	300 μm–30 μm	4×10^{-3} eV–4×10^{-2} eV
Mid-infrared (MIR)	10^{13} Hz–6×10^{13} Hz	30 μm–5 μm	4×10^{-2} eV–0.3 eV
Near-infrared (NIR)	6×10^{13} Hz–3×10^{14} Hz	5 μm–1 μm	0.3 eV–1 eV
Optical*	3×10^{14} Hz–1×10^{15} Hz	1000 nm–300 nm	1 eV–4 eV
Ultraviolet	1×10^{15} Hz–3×10^{16} Hz	300 nm–10 nm	4 eV–100 eV
Soft x-rays	3×10^{16} Hz–3×10^{18} Hz	10 nm–0.1 nm	100 eV–10 keV
Hard x-rays	3×10^{18} Hz–3×10^{20} Hz	0.1 nm–0.001 nm	10 keV–1 MeV
Gamma rays	3×10^{20} Hz–3×10^{25} Hz	10^{-12} m–10^{-17} m	1 MeV–100 GeV
TeV	3×10^{25} Hz–3×10^{27} Hz	10^{-17} m–10^{-20} m	0.1 TeV–30 TeV

* The visible part of the optical spectrum extends roughly from 800 to 400 nm. Optical wavelengths are also often given in Ångstrom: 1 nm = 10 Å.

Index