Chapter Seven

Supermassive Black Holes

7.1 Quasars and Black Holes: An Overview

A black hole is the end product of the complete gravitational collapse of a material object, such as a massive star. It is surrounded by a horizon from which even light cannot escape. Black holes have the dual virtues of being extraordinarily simple solutions to Einstein's equations of gravity (as they are characterized only by their mass, charge, and spin) but also the most disparate from their Newtonian analogs. In Einstein's theory of gravity, black holes represent the ultimate prisons: you can check in, but you can never check out.

Ironically, black hole environments are the brightest objects in the universe. Of course, it is not the black hole that is shining but, rather, the surrounding gas heated by viscously rubbing against itself and shining as it spirals into the black hole like water going down a drain, never to be seen again. The origin of the radiated energy is the release of gravitational binding energy as the gas falls into the deep gravitational potential well of the black hole. More than 10 percent of the mass of the accreting material can be converted into heat (more than an order of magnitude beyond the maximum efficiency of nuclear fusion). Astrophysical black holes appear in two flavors: stellar-mass black holes that form when massive stars die, and the monstrous supermassive black holes that sit at the center of galaxies, reaching masses of up to 10 billion Suns. The latter type are observed as *active galactic nuclei* (AGN). It is by studying these accreting black holes that all our observational knowledge of black holes has been obtained.

A quasar—the most powerful type of AGN—is a point-like ("quasi-stellar") bright source at the center of a galaxy. Many lines of evidence indicate that a quasar involves a supermassive black hole that is accreting gas from the core of its host galaxy. The supply of large quantities of fresh gas is often triggered by a merger between two galaxies. The infalling gas heats up as it spirals toward the black hole and dissipates its rotational energy through viscosity. The gas is expected to drift inward in an accretion disk until it reaches the last possible stable orbit according to general relativity. Interior to this point, the gas plunges into the black hole in such a short time that it has no opportunity to radiate most of its thermal energy. However, as is described in detail in §7.2, the fraction of the rest mass of the gas that gets radiated away just outside this orbit is high, ranging between 5.7% for a nonspinning black hole to 42.3% for a maximally spinning black hole (see Figure 7.5). This "radiative efficiency" is far greater

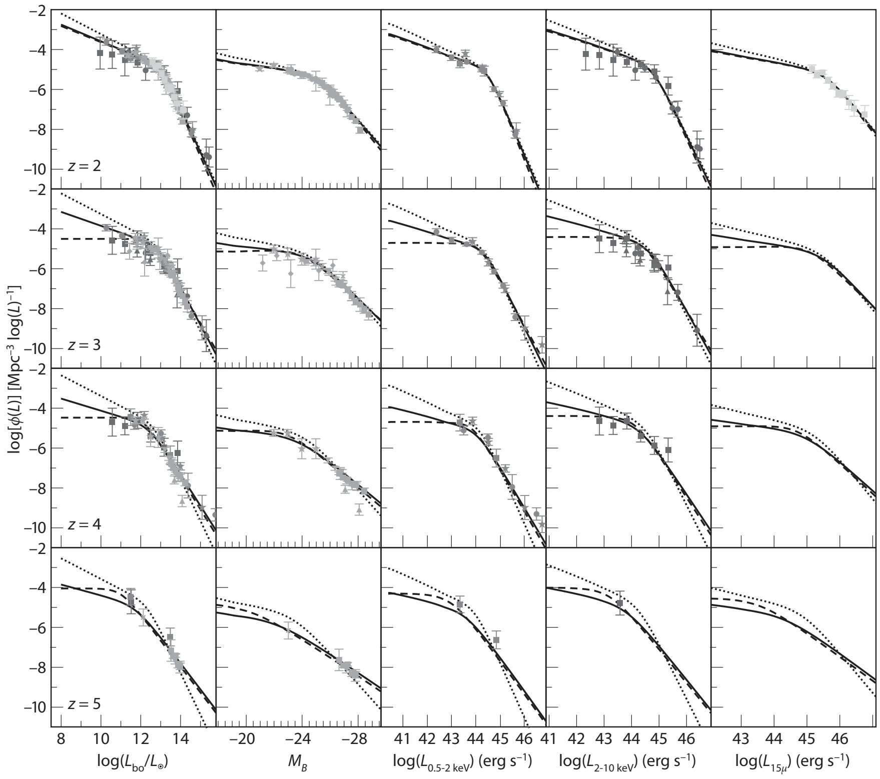

Figure 7.1 Redshift evolution of the luminosity function of quasars at different observed wavelengths: B-band (center-left panels), soft X-rays (0.5–2 keV) (center), hard X-rays (2–10 keV) (center-right), and mid-IR (15 μm) (right). The left panels show the distribution of bolometric luminosities (integrated over all wavelengths). Lines show the best-fit evolving double power-law model to data points at all redshifts (solid), the best-fit model at the given redshift (dashed), and the best-fit model that allows only the break luminosity to evolve (dotted). Hopkins, P. F., Richards, G. T., & Hernquist, L., *Astrophys. J.* **654**, 731 (2007). Reproduced by permission of the AAS.

than the mass–energy conversion efficiency provided by nuclear fusion in stars, which is $< 0.7\%$.

Fortunately, quasars are very easy to see when the accretion occurs through a thin disk, and we have a great deal of demographic information on their properties out to very high redshifts. Quasar activity is observed in a small fraction of all galaxies at any cosmic epoch. Figure 7.1 shows the evolution of the luminosity function of quasars at different observed wavelengths in the redshift interval $z = 2$–5. Mammoth black holes weighing more than a billion solar masses have been discovered at redshifts as high as $z \sim 6$–7, less than a billion years after the Big Bang. The highest redshift quasar known (as of winter 2012) is ULAS J1120+0641 at $z = 7.085$ (only 0.77 Gyr after the Big Bang), with a bolometric luminosity of $6.3 \times 10^{13}\ L_\odot$ and an estimated black hole mass of $2 \times 10^9\ M_\odot$;

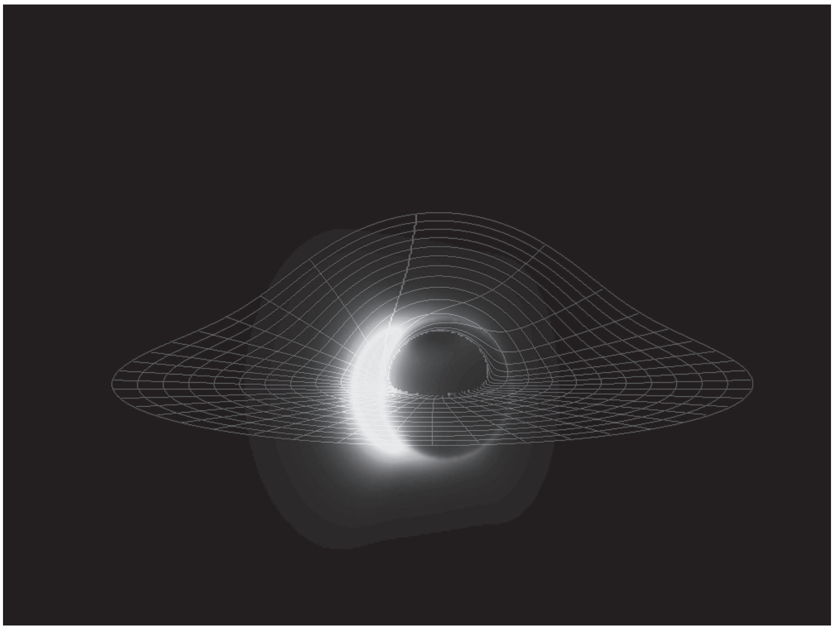

Figure 7.2 Simulated image of an accretion flow around a black hole spinning at half its maximum rate, from a viewing angle of 10° relative to the rotation axis (see *Color Plate 12* for a color version of this figure). The coordinate grid in the equatorial plane of the spiraling flow shows how strong lensing around the black hole bends the back of the apparent disk up. The left side of the image is brighter owing to its rotational motion toward the observer. The bright arcs are generated by gravitational lensing. A dark silhouette appears around the location of the black hole because the light emitted by gas behind it disappears into the horizon and cannot be seen by an observer on the other side. Recently, the technology for observing such an image from the supermassive black holes at the centers of the Milky Way and M87 galaxies has been demonstrated as feasible [Doeleman, S., et al., *Nature* **455**, 78 (2008)]. To obtain the required resolution of tens of micro-arcseconds, interferometers operating at millimeter wavelengths across the earth are necessary. Broderick, A., & Loeb, A. *Journal of Physics Conf. Ser.* **54**, 448 (2006); *Astrophys. J.* **697**, 1164 (2009). Reproduced by permission of the AAS.

we show a partial spectrum of this quasar in Figure 11.8, where we discuss this object in the context of the reionization of the Universe.

There is clear and direct evidence for supermassive black holes, even beyond the AGN population. In our own Milky Way galaxy, stars are observed to zoom around the Galactic center at speeds of up to 10,000 $\mathrm{km\,s^{-1}}$, owing to the strong gravitational acceleration near the central black hole (with a mass $\sim 4 \times 10^6\ M_\odot$).[1] But closer-in observations are forthcoming. Existing technology should soon be able to image the silhouette of the supermassive black holes in the Milky Way and M87 galaxies directly (see Figure 7.2).

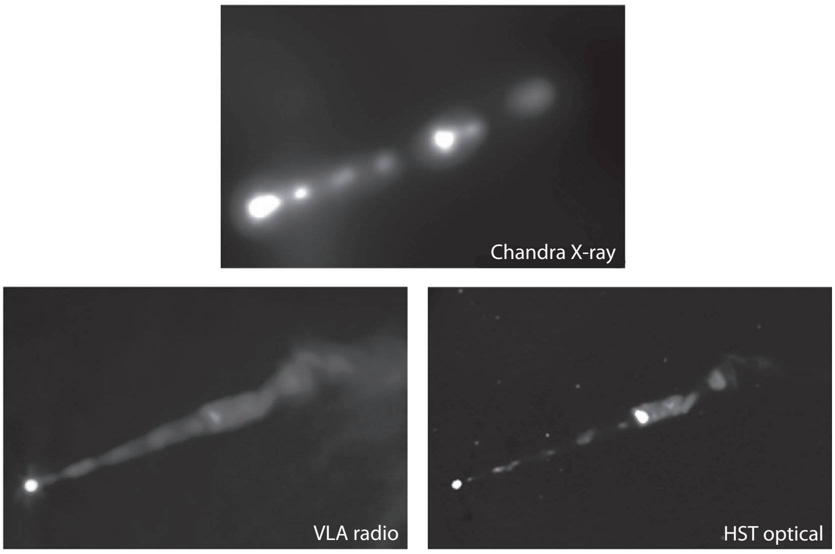

Figure 7.3 Multiwavelength images of the highly collimated jet emanating from the supermassive black hole at the center of the giant elliptical galaxy M87 (see *Color Plate 13* for a color version of this figure). The X-ray image (*top*) was obtained with the Chandra X-ray observatory, the radio image (*bottom left*) was obtained with the Very Large Array (VLA), and the optical image (*bottom right*) was obtained with the Hubble Space Telescope (HST). Courtesy of CXO/NASA.

Nevertheless, many questions remain about black holes. If the accreting material is organized into a thin disk, where the gas can efficiently radiate its released binding energy, then its theoretical modeling is straightforward. Less well understood are radiatively inefficient accretion flows, in which the inflowing gas obtains a thick geometry. It is generally unclear how gas migrates from large radii to near the horizon and how, precisely, it falls into the black hole. For example, we presently have very poor constraints on how magnetic fields embedded and created by the accretion flow are structured, and how that structure affects the observed properties of astrophysical black holes. While it is beginning to be possible to perform computer simulations of the entire accreting region, we are decades away from true ab initio calculations, and thus observational input plays a crucial role in deciding between existing models and motivating new ideas.

More embarrassing is our crude understanding of black hole jets (see Figure 7.3) and other feedback mechanisms. These extraordinary exhibitions of the power of black holes are moving at nearly the speed of light and involve narrowly collimated outflows whose base has a size comparable to the solar system, while their front reaches scales comparable to the distance between galaxies.[2] Unresolved issues are as basic as what jets are made of (whether electrons and protons, electrons and positrons, or primarily electromagnetic fields) and how they are accelerated in the first place. The answers to these questions depend

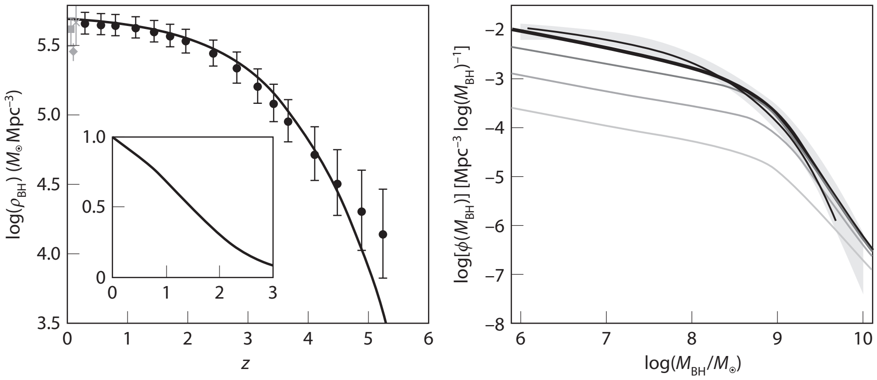

Figure 7.4 *Left panel:* The comoving black hole mass density of quasars from the data (circles) and the best-fit luminosity function (solid line). The inset shows the fraction of the mass density relative to today's value, $\rho_{\mathrm{BH}}(z)/\rho_{\mathrm{BH}}(0)$, on a linear scale. *Right:* The black hole mass function at $z = 0$ (thick black line) and $z = 1, 2, 3$ (from top to bottom). The shaded region shows the 1σ observational uncertainty. Hopkins, P. F., Richards, G. T., & Hernquist, L., *Astrophys. J.* **654**, 731 (2007). Reproduced by permission of the AAS.

on critically on the role of the black hole spin in the jet-launching process. The bright radiation and powerful jets of AGN can have dramatic feedback effects on the black hole's host galaxy.

If massive black holes grew at early cosmic times, should their remnants be around us today? Indeed, searches for black holes in local galaxies have found that every galaxy with a stellar spheroid harbors a supermassive black hole at its center, which implies that quasars are rare simply because their activity is short-lived. The inferred growth in the comoving mass function of black holes along with its integral over all black hole masses (i.e., the comoving mass density) are shown in Figure 7.4. Moreover, there appears to be a tight correlation between black hole mass and the gravitational potential-well depth of their host spheroids of stars (as measured by the velocity dispersion of these stars), as described in detail in §7.5.1.

This suggests that the black holes grow up to the point where the heat they deposit into their environment or the piston effect from their winds prevents additional gas from feeding them further, or else the feedback from star formation in the vicinity of the black holes affects or controls their self-regulation. The black hole is like a baby who gets more energetic as he eats more at the dinner table, until his hyperactivity becomes so intense that he pushes the food off the table and cannot eat any more. This *principle of self-regulation* explains why quasars are short-lived and why the final black hole mass is dictated by the depth of the potential in which the gas feeding it resides.[3] Evidently, the growth of supermassive black holes is intimately linked to the hierarchical growth of their host galaxies. Most black holes today are dormant or "starved" because the gas around them was mostly used up in making the stars, or because their activity heated or pushed it away a long time ago.

This state of affairs can easily be understood from the fact that the binding energy per unit mass in a typical galaxy corresponds to velocities v of hundreds of kilometers per second or a fraction $\sim (v/c)^2 \sim 10^{-6}$ of the binding energy per unit mass near a black hole. Hence, a small amount of gas that releases its binding energy near a black hole can have a large effect on the rest of the gas in the galaxy.

Why did the collapsed matter in the Universe end up making galaxies and not black holes? One would have naively expected spherical collapse to end with the formation of a point mass at its center. But, as it turns out, tides from neighboring objects torque the infalling material and induce asphericity and some spin into the final collapse. The induced angular momentum prevents the gas from reaching the center on a direct plunging orbit. After the gas cools and loses its pressure support against gravity, it instead assembles into a disk in which the centrifugal force balances gravity. The finite size of the luminous region of galaxies is then dictated by the characteristic spin acquired by galaxy halos, which typically corresponds to a rotational velocity that is $\sim$5% of the virial circular velocity, with a negligible dependence on halo mass. In the previous two chapters we discussed how this material forms stars and, eventually, galaxies.

But star formation does not imply that no gas accumulates at the center. In fact, galactic spheroids are observed generically to harbor a central black hole, whose formation is most likely linked to the small mass fraction of galactic gas ($<$0.1%) with unusually small amount of angular momentum. The small mass fraction of the central black holes implies that their gravitational effect is restricted to the innermost cusp of their host galaxy.

In this chapter, we study formation mechanisms for supermassive black holes, their observable characteristics, and their interactions with their host galaxies and the wider Universe. We begin with a short introduction to the properties of black holes in general relativity.

7.2 Basic Principles of Astrophysical Black Holes

In Newtonian gravity, the gravitational field at any radius outside a spherical mass distribution depends only on the mass interior to that radius. This result is also true in Einstein's general relativity, where Birkhoff's theorem (see §1.2.2) states that the only vacuum, spherically symmetric gravitational field is that described by the static *Schwarzschild metric,*

$$ds^2 = -\left(1 - \frac{r_{\rm Sch}}{r}\right) c^2 dt^2 + \left(1 - \frac{r_{\rm Sch}}{r}\right)^{-1} dr^2 + r^2 d\Omega, \tag{7.1}$$

where $d\Omega = (d\theta^2 + \sin^2\theta\, d\phi^2)$. The *Schwarzschild radius* is related to the mass M of the central (nonspinning) black hole,

$$r_{\rm Sch} = \frac{2GM}{c^2} = 2.95 \times 10^5 \text{ cm} \left(\frac{M}{1\, M_\odot}\right). \tag{7.2}$$

The black hole horizon, $r_{\rm Hor}$ ($= r_{\rm Sch}$ here), is a spherical boundary from which no particle can escape. [The coordinate singularity of the Schwarzschild metric at $r = r_{\rm Sch}$ can be removed through a transformation to the *Kruskal* coordinate system $(r, t) \rightarrow (u, v)$, where $u = (r/r_{\rm Sch} - 1)^{1/2}\, e^{r/2r_{\rm Sch}} \cosh(ct/2r_{\rm Sch})$; $v = u \tanh(ct/2r_{\rm Sch})$.] The existence of a spatial region into which particles may fall but never come out breaks the time-reversal symmetry that characterizes the equations of quantum mechanics. Any grander theory that would unify quantum mechanics and gravity must remedy this conceptual inconsistency.

In addition to its mass M, a black hole can be characterized only by its spin J and electric charge Q (similarly to an elementary particle). In astrophysical circumstances, any initial charge of the black hole would quickly be neutralized through the polarization of the background plasma and the preferential infall of electrons or protons. The residual electric charge would exert an electric force on an electron comparable to the gravitational force on a proton, $eQ \sim GMm_p$, which implies that $(Q^2/GM^2) \sim Gm_p^2/e^2 \sim 10^{-36}$ and a negligible contribution of the charge to the metric. A spin, however, may modify the metric considerably.

The general solution of Einstein's equations for a spinning black hole was derived by Kerr in 1963 and can be written most conveniently in Boyer-Lindquist coordinates:

$$ds^2 = -\left(1 - \frac{r_{\rm Sch} r}{\Sigma_k}\right) c^2\, dt^2 - \frac{2 j r_{\rm Sch} r \sin^2\theta}{\Sigma_k} c dt\, d\phi + \frac{\Sigma_k}{\Delta} dr^2$$
$$+ \Sigma_k\, d\theta^2 + \left(r^2 + j^2 + \frac{r_{\rm Sch} j^2 r \sin^2\theta}{\Sigma_k}\right) \sin^2\theta\, d\phi^2, \tag{7.3}$$

where the black hole is rotating in the ϕ direction, $j = [J/Mc]$ is the normalized angular momentum per unit mass (in units of cm), $\Delta = r^2 - r r_{\rm Sch} + j^2$, and $\Sigma_k = r^2 + j^2 \cos^2\theta$. The dimensionless ratio $a = j/(GM/c^2)$ is bounded by unity, and $a = 1$ corresponds to a maximally rotating black hole. The horizon radius $r_{\rm Hor}$ is now located at the larger root of the equation $\Delta = 0$, namely, $r_+ = (1/2) r_{\rm Sch}[1 + (1 - a^2)^{1/2}]$. The Kerr metric converges to the Schwarzschild metric for $a = 0$. There is no Birkhoff's theorem for a rotating black hole.

Test particle orbits around black holes can be simply described in terms of an effective potential. For photons around a Schwarzschild black hole, the potential is simply $V_{\rm ph} = (1 - r_{\rm Sch}/r)/r^2$, which leads to circular photon orbits at a radius $r_{\rm ph} = (3/2) r_{\rm Sch}$. For a spinning black hole,

$$r_{\rm ph} = r_{\rm Sch}\left[1 + \cos\left(\frac{2}{3}\cos^{-1}[\pm a]\right)\right], \tag{7.4}$$

where the upper sign refers to orbits that rotate in the direction opposite that of the black hole (retrograde orbits), and the lower sign to corotating (prograde) orbits. For a maximally rotating black hole ($|a| = 1$), the photon orbit radius is $r_{\rm ph} = (1/2) r_{\rm Sch}$ for a prograde orbit and $2 r_{\rm Sch}$ for a retrograde orbit.

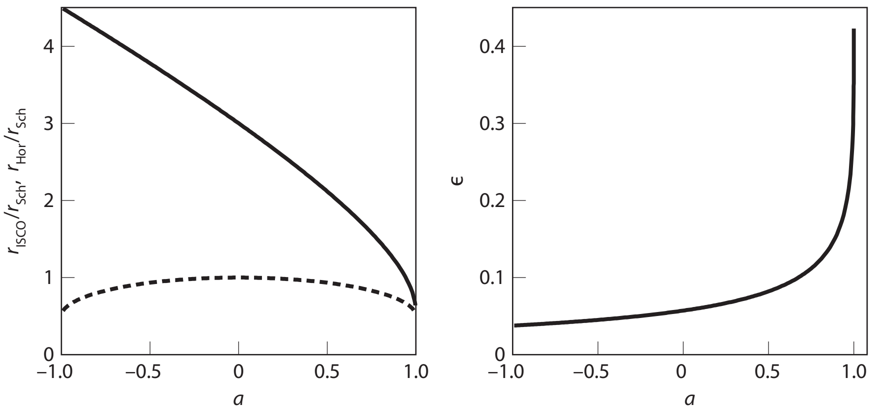

Figure 7.5 The left panel shows the radius of the black hole horizon r_{Hor} (dashed line) and the innermost stable circular orbit (ISCO) around it r_{ISCO} (solid line), in units of the Schwarzschild radius r_{Sch} (see equation 7.2), as functions of the black hole spin parameter a. The limiting value of $a = 1$ ($a = -1$) corresponds to a corotating (counterrotating) orbit around a maximally spinning black hole. The binding energy of a test particle at the ISCO determines the radiative efficiency ϵ of a thin accretion disk around the black hole, shown in the right panel.

Circular orbits of massive particles exist when the first derivative of their effective potential (including angular momentum) with respect to radius vanishes, and these orbits are stable if the second derivative of the potential is positive. The radius of the *innermost stable circular orbit (ISCO)* defines the inner edge of any disk of particles in circular motion (such as fluid elements in an accretion disk). At smaller radii, gravitationally bound particles plunge into the black hole on a dynamical time. This radius of the ISCO is given by[4]

$$r_{\mathrm{ISCO}} = \frac{1}{2} r_{\mathrm{Sch}} \left\{ 3 + Z_2 \pm [(3 - Z_1)(3 + Z_1 + 2Z_2)]^{1/2} \right\}, \tag{7.5}$$

where $Z_1 = 1+(1-a^2)^{1/3}[(1+a)^{1/3}+(1-a)^{1/3}]$, and $Z_2 = (3a^2+Z_1^2)^{1/2}$. Figure 7.5 shows the radius of the ISCO as a function of spin. The binding energy of particles at the ISCO defines their maximum radiative efficiency, because they spend a short time on their plunging orbit interior to the ISCO. This efficiency is given by

$$\epsilon = 1 - \frac{r^2 - r_{\mathrm{Sch}} r \mp j\sqrt{\frac{1}{2} r_{\mathrm{Sch}} r}}{r(r^2 - \frac{3}{2} r_{\mathrm{Sch}} r \mp 2j\sqrt{\frac{1}{2} r_{\mathrm{Sch}} r})^{1/2}}. \tag{7.6}$$

The efficiency ranges between $\epsilon = (1 - \sqrt{8/9}) = 5.72\%$ for $a = 0$, to $(1 - \sqrt{1/3}) = 42.3\%$ for a prograde (corotating) orbit with $a = 1$, and $(1-\sqrt{25/27}) = 3.77\%$ for a retrograde orbit.

7.3 Accretion of Gas onto Black Holes

7.3.1 Bondi Accretion

Consider a black hole embedded in a hydrogen plasma of uniform density $\rho_0 = m_p n_{\rm H}$ and temperature T_0. The thermal protons in the gas are moving around at roughly the sound speed $c_s \sim \sqrt{k_B T/m_p}$. The black hole gravity can drive accretion of gas particles that are gravitationally bound to it, namely, interior to the radius of influence, $r_{\rm inf} \sim GM/c_s^2$. The steady mass flux of particles entering this radius is $\rho_0 c_s$. Multiplying this flux by the surface area associated with the radius of influence gives the supply rate of fresh gas,

$$\dot{M} \approx \pi r_{\rm inf}^2 \rho_0 c_s = 15 \left(\frac{M}{10^8\ M_\odot}\right)^2 \left(\frac{n_{\rm H}}{1\ {\rm cm}^{-3}}\right) \left(\frac{T_0}{10^4\ {\rm K}}\right)^{-3/2} M_\odot\ {\rm yr}^{-1}. \qquad (7.7)$$

In a steady state this supply rate equals the mass accretion rate into the black hole (see also §5.2.1).

The explicit steady-state solution to the conservation equations of the gas (mass, momentum, and energy), which is self-similar, was first derived by Bondi in 1952. This exact solution introduces a correction factor of the order of unity to equation (7.7).[5] Well inside the sonic radius (i.e., the point at which the infall and sound speeds are equal), the velocity is close to freefall $u \sim (2GM/r)^{1/2}$, and the gas density is $\rho \sim \rho_0 (r/r_{\rm inf})^{-3/2}$. The radiative efficiency is small, because either the gas is tenuous, so that its cooling time is longer than its accretion (free-fall) time, or the gas is dense, and the diffusion time of the radiation outward is much longer than the free-fall time. If the inflowing gas contains near-equipartition magnetic fields, then cooling through synchrotron emission typically predominates over free-free emission.

A black hole that is moving with a velocity V relative to a uniform medium accretes at a lower rate than a stationary black hole. At high velocities, the radius of influence of the black hole is then $\sim GM/V^2$, which suggests that the sound speed c_s should be crudely replaced with $\sim (c_s^2 + V^2)^{1/2}$ in equation (7.7). A similar suppression factor applies for the accretion of baryons onto dark matter halos when the baryons have a net bulk velocity relative to the dark matter (see §2.1.2 and 3.2.2).

7.3.2 Thin-Disk Accretion

If the inflow is endowed with rotation, the gas will reach a centrifugal barrier from which it can accrete farther inward only after its angular momentum is transported away. This limitation follows from the steeper radial scaling of the centrifugal acceleration ($\propto r^{-3}$ if angular momentum is conserved; see equation 5.33) compared with the gravitational acceleration ($\propto r^{-2}$). Near the centrifugal barrier, where the gas is held against gravity by rotation, an accretion disk will form around the black hole, centered on the plane perpendicular to the rotation axis. The accretion time will then be dictated by the rate at which angular momentum is transported through viscous stress, and it could be significantly longer than the free-fall time for a nonrotating flow (as in the solution described by the Bondi accretion model). As the gas settles to a disk, the dissipation of its kinetic energy into heat will make the disk thick and hot, with a proton temperature close to the gravitational potential energy per proton, $\sim 10^{12}$ K$(r/r_{\rm Sch})^{-1}$ in the absence of radiative processes. However, if the cooling time of the gas is shorter than the viscous time, then a thin disk will form. This latter regime is realized for the high gas inflow rate during the processes (such as galaxy mergers) that feed quasars. To better understand such objects, we start by exploring the structure of thin disks that characterize the high accretion rate of quasars.[6]

We imagine a planar thin disk of cold gas orbiting a central black hole and wish to describe its structure in polar coordinates (r, ϕ). Each gas element orbits at the local Keplerian velocity $v_\phi = r\Omega = (GM/r)^{1/2}$, where Ω is the angular velocity, and spirals slowly inward with radial velocity $v_r \ll v_\phi$ as viscous torques transport its angular momentum to the outer part of the disk. The associated viscous stress generates heat, which is radiated away locally from the the disk surface. We assume that the disk is fed steadily, and so it manifests a constant mass accretion rate at all radii. Mass conservation implies that

$$\dot{M} = 2\pi r \Sigma v_r = \text{constant}, \tag{7.8}$$

where $\Sigma(r)$ is the surface mass density of the disk.

In the limit of a geometrically thin disk with a scale height $h \ll r$, the hydrodynamic equations decouple in the radial and vertical directions. We start with the radial direction. The Keplerian velocity profile introduces shear, which dissipates heat as neighboring fluid elements rub against one another. The concept of shear viscosity can easily be understood in the one-dimensional example of a uniform gas whose velocity along the y-axis varies linearly with the x-coordinate, $V = V_0 + (dV_y/dx)x$. A gas particle moving at the typical thermal speed v traverses a mean free path λ along the x-axis before it collides with other particles and shares its y-momentum with them. The y-velocity is different across a distance λ by an amount $\Delta V \sim \lambda\, dV_y/dx$. Since the flux of particles streaming along the x-axis is $\sim nv$, where n is the gas density, the net flux of y-momentum transported per unit time, $\sim nvm\Delta V$, is linear in the velocity gradient $\eta\, dV_y/dx$, with a viscosity coefficient $\eta \sim \rho v \lambda$ (in g cm^{-1} s^{-1}), where $\rho = mn$ is the mass density of the gas. Since the excess kinetic energy density across a mean free path, $(1/2)\rho(\lambda\, dV_y/dx)^2$, is dissipated every collision time $\sim (\lambda/v)$, viscosity heats the gas at a rate per unit volume of $\mathcal{H}_{\rm vis} \sim [\eta(dV_y/dx)]^2/\eta$.

Similar arguments show that within a Keplerian accretion disk, the flux of ϕ-momentum transported in the positive r-direction is given by the viscous stress $f_\phi = (3/2)\eta\Omega$. This dissipation is expected to be effective down to the ISCO, from which the gas plunges into the black hole over a free-fall time. We therefore set the inner boundary of the disk as $r_{\rm ISCO}$, depicted in Figure 7.5. Conservation of angular momentum requires that the net rate of its change within a radius r be equal to the viscous torque, namely,

$$f_\phi \times (2\pi r \times 2h) \times r = \dot{M}\left[(GMr)^{1/2} - (GMr_{\rm ISCO})^{1/2}\right]. \tag{7.9}$$

The production rate of heat per unit volume by the viscous stress is given by $\mathcal{H}_{\rm vis} = f_\phi^2/\eta$. Substituting for f_ϕ in equation (7.9), we have

$$2h\mathcal{H}_{\rm vis} = \frac{3\dot{M}}{4\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.10}$$

This power provides a local radiative flux that leaves the system vertically from the top and bottom surfaces of the disk,

$$F = \frac{1}{2} \times 2h\mathcal{H}_{\rm vis} = \frac{3\dot{M}}{8\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.11}$$

The total luminosity of the disk is given by

$$L = \int_{r_{\rm ISCO}}^{\infty} 2F \times 2\pi r\, dr = \frac{1}{2}\frac{GM\dot{M}}{r_{\rm ISCO}}, \tag{7.12}$$

where we have ignored general-relativistic corrections to the dynamics of the gas and the propagation of the radiation it emits.

In the absence of any vertical motion, momentum balance in the vertical z-direction yields

$$\frac{1}{\rho}\frac{dp}{dz} = -\frac{GM}{r^2}\frac{z}{r}, \tag{7.13}$$

where $z \ll r$ and p and ρ are the gas pressure and density, respectively. This equation fixes the disk scale height $h \approx c_s/\Omega$, where $c_s \approx (P/\rho)^{1/2}$ is the sound speed.

Because of the short mean free path for particle collisions, the particle-level viscosity is negligible in accretion disks. However, such disks are susceptible to the powerful *magnetorotational instability* (MRI) that amplifies magnetic turbulence over an orbital time. The origin of the instability can easily be understood by imagining two fluid elements that are threaded by a single magnetic field line and are slightly displaced from each other in the radial direction. The magnetic field acts as a spring owing to its tension. In a Keplerian disk the inner fluid element orbits more rapidly than the outer element, causing the spring to stretch. The inner fluid element is then forced by the spring to slow down, to reduce its angular momentum, and therefore to move to a lower orbit. The outer fluid element, meanwhile, is forced by the spring to speed up, to increase its angular momentum, and therefore to move to a higher orbit. The spring tension increases as the two fluid elements separate farther, and eventually the process runs away. The magnetorotational instability[7] is likely to develop turbulent eddies in the disk that are much more effective at transporting its angular momentum than particle viscosity. In this case λ and v should be replaced by the typical size and velocity of an eddy. The largest value that these variables can attain are the scale height h and sound speed c_s in the disk, which implies that $f_\phi < (\rho c_s h)\Omega \approx \rho c_s^2 \approx p$. We may then parameterize the viscous stress as some fraction α of its maximum value, $f_\phi = \alpha p$.

The total pressure p in the disk is the sum of the gas pressure $p_{\rm gas} = 2(\rho/m_p)k_B T$, and the radiation pressure, $p_{\rm rad} = 1/3 a_{\rm rad} T^4$. We define the fractional contribution of the gas to the total pressure as

$$\beta \equiv \frac{p_{\rm gas}}{p}. \tag{7.14}$$

In principle, the viscous stress may be limited by the gas pressure only; to reflect this possibility, we write $f_\phi = \alpha p \beta^b$, where b is 0 or 1 if the viscosity scales with the total pressure or just the gas pressure, respectively.

Since the energy of each photon is just its momentum times the speed of light, the radiative energy flux is simply given by the change in the radiation pressure (momentum flux) per photon mean free path,

$$F = -c\,\frac{dp_{\rm rad}}{d\tau}, \tag{7.15}$$

where the optical depth τ is related to the frequency-averaged (so-called Rosseland-mean) opacity coefficient of the gas, κ,

$$\tau = \int_0^h \kappa\rho\, dz \approx \frac{1}{2}\kappa\Sigma, \tag{7.16}$$

where $\Sigma = 2h\rho$. For the characteristic mass density ρ and temperature T encountered at the midplane of accretion disks around supermassive black holes, there are two primary sources of opacity: *electron scattering* with

$$\kappa_{\rm es} = \frac{\sigma_{\rm T}}{m_p} = 0.4\ {\rm cm^2\,g^{-1}}, \tag{7.17}$$

and *free-free* absorption with

$$\kappa_{\rm ff} \approx 8\times 10^{22}{\rm cm^2\,g^{-1}}\left(\frac{\rho}{\rm g\,cm^{-3}}\right)\left(\frac{T}{\rm K}\right)^{-7/2}, \tag{7.18}$$

where we have assumed a pure hydrogen plasma for simplicity.

It is customary to normalize the accretion rate $\dot{M}$ in the disk relative to the so-called Eddington rate $\dot{M}_E$, which produces the maximum possible disk luminosity, L_E (see the derivation in equation 7.33). When the luminosity approaches the Eddington limit, the disk bloats and h approaches r, violating the thin-disk assumption. We write $\dot{m} = (\dot{M}/\dot{M}_E)$, with $\dot{M}_E \equiv (L_E/\epsilon c^2)$, where ϵ is the radiative efficiency for converting rest mass to radiation near the ISCO (shown in Figure 7.5).

Utilizing the preceding equations, we are now in a position to derive the scaling laws that govern the structure of the disk far away from the ISCO. For this purpose we use the following dimensionless parameters: $r_1 = (r/10r_{\rm Sch})$, $M_8 = (M/10^8\ M_\odot)$, $\dot{m}_{-1} = (\dot{m}/0.1)$, $\alpha_{-1} = (\alpha/0.1)$ and $\epsilon_{-1} = (\epsilon/0.1)$.

In local thermodynamic equilibrium, the emergent flux from the surface of the disk (equation 7.11) can be written in terms of the midplane disk temperature T as $F \approx ca_{\rm rad}T^4/\kappa\Sigma$. The surface temperature of the disk is then roughly

$$T_d \approx \left(\frac{4F}{a_{\rm rad}}\right)^{1/4} = 10^5\ {\rm K}\ M_8^{-1/4}\dot{m}_{-1}^{1/4}r_1^{-3/4}\left[1-\left(\frac{r}{r_{\rm ISCO}}\right)^{1/2}\right]. \tag{7.19}$$

Note that the disk surface temperature increases at small black hole masses and reaches the X-ray regime for stellar-mass black holes. (Nonthermal X-ray emission from a hot corona or a jet can supplement this disk emission.) Stellar-mass black holes can therefore be important X-ray sources at high redshifts, especially if they get incorporated into a binary system where they accrete gas from a companion star. In the local Universe, black hole X-ray binaries come in two flavors, depending on the mass of the companion star: *low-mass X-ray binaries*, in which a low-mass companion transfers mass owing to the tidal force exerted by the black hole, and *high-mass X-ray binaries (BH-HMXBs* or *microquasars)*, in which the companion is a massive star that can also transfer mass to the black hole through a wind. At redshifts $z > 6$, when the age of the Universe was short, BH-HMXBs were probably most important, since they are known to produce their X-rays over a short lifetime ($< 10^9$ yr). The cumulative X-ray emission from BH-HMXBs is expected to be proportional to the star formation rate. If, indeed, the early population of stars was tilted toward high masses, and binaries were common, BH-HMXBs may have been more abundant per star formation rate in high-redshift galaxies. As we discuss elsewhere, the X-rays produced by BH-HMXBs may have had important observable effects as they catalyzed H_2 formation (§6.2), heated the IGM (§9.8.2), and modified the 21-cm signal from neutral hydrogen (§12.3.2). Their overall influence was, however, limited: hydrogen could not have been reionized by X-ray sources based on current limits on the unresolved component of the X-ray background. Throughout this chapter, we focus our attention on supermassive black holes, which are brighter and hence easier to detect individually at high redshifts.

For supermassive black holes, the accretion disk can be divided radially into three distinct regions,[8]

1. *Inner region:* where radiation pressure and electron-scattering opacity dominate
2. *Middle region:* where gas pressure and electron-scattering opacity dominate
3. *Outer region:* where gas pressure and free-free opacity dominate

The boundary between regions 1 and 2 is located at the radius

$$r_{1,\rm im} \approx 54\,\alpha_{-1}^{2/21}(\dot{m}_{-1}/\epsilon_{-1})^{16/21}M_8^{2/21} \quad \text{if } b = 1, \tag{7.20}$$

$$58\,\alpha_{-1}^{2/21}(\dot{m}_{-1}/\epsilon_{-1})^{16/21}M_8^{2/21} \quad \text{if } b = 0, \tag{7.21}$$

and the transition radius between regions 2 and 3 is

$$r_{1,\rm mo} \approx 4\times10^2\,(\dot{m}_{-1}/\epsilon_{-1})^{2/3}. \tag{7.22}$$

The surface density and scale height of the disk are given by

Inner region:

$$\Sigma(r) \approx (3 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5} \quad \text{if } b = 1, \tag{7.23}$$

$$(8 \times 10^2\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-1}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{-1} r_1^{3/2} \quad \text{if } b = 0, \tag{7.24}$$

$$h(r) \approx r_{\rm Sch}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right). \tag{7.25}$$

Middle region:

$$\Sigma(r) \approx (3 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5}, \tag{7.26}$$

$$h(r) \approx 1.4 \times 10^{-2} r_{\rm Sch}\alpha_{-1}^{-1/10}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{1/5} M_8^{-1/10} r_1^{21/20}. \tag{7.27}$$

Outer region:

$$\Sigma(r) \approx (6 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{0.1}}\right)^{7/10} M_8^{1/5} r_1^{-3/4}, \tag{7.28}$$

$$h(r) \approx 10^{-2} r_{\rm Sch}\alpha_{-1}^{-1/10}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/20} M_8^{-1/10} r_1^{9/8}. \tag{7.29}$$

The midplane temperature is given by

$$T_m(r) \approx \left(16\pi^2\right)^{-1/5}\left(\frac{m_p}{k_B\sigma_T}\right)^{1/5} \alpha^{-1/5}\kappa^{1/5}\dot{M}^{2/5}\Omega^{3/5}\beta^{-(1/5)(b-1)}. \tag{7.30}$$

The preceding scaling laws ignore the self-gravity of the disk. This assumption is violated at large radii. The instability of the disk to gravitational fragmentation due to its self-gravity occurs when the so-called Toomre parameter, $Q = (c_s\Omega/\pi G\Sigma)$, drops below unity (see §5.2.3). For the preceding scaling laws of the outer disk, the onset of fragementation occurs at the outer radius,

$$r_{1,\rm out} \approx 2 \times 10^4 \alpha_{-1}^{28/45}\,(\dot{m}_{-1}/\epsilon_{-1})^{-22/45}\,M_8^{52/45}. \tag{7.31}$$

Outside this radius, the disk gas fragments into stars, and the stars may migrate inward as the gas accretes onto the black hole. The energy output from stellar winds and supernovae will supplement the viscous heating of the disk and may regulate the disk to have $Q \sim 1$ outside this boundary. We therefore conclude that star formation will inevitably occur on larger scales, before the gas is driven into the accretion disk that feeds the central black hole. Indeed, the broad emission lines of quasars display very high abundances of heavy elements in the spectra out to arbitrarily high redshifts. Since the total amount of mass in the disk interior to this radius constitutes only a small fraction of the mass of the supermassive black hole, quasars must be fed by gas that crosses this boundary after being vulnerable to fragmentation.[9]

7.3.3 Radiatively Inefficient Accretion Flows

When the accretion rate is considerably lower than its Eddington limit ($\dot{M}/\dot{M}_E < 10^{-2}$), the gas inflow switches to a different mode, called a *radiatively inefficient accretion flow* (RIAF) or an *advection-dominated accretion flow* (ADAF), in which either the cooling time or the photon diffusion time is much longer than the accretion time of the gas, and heat is mostly advected with the gas into the black hole. At the low gas densities and high temperatures characterizing this accretion mode, the Coulomb coupling is weak, and the electrons do not heat up to the proton temperature, even with the aid of plasma instabilities. Viscosity heats primarily the protons, since they carry most of the momentum. The other major heat source, compression of the gas, also heats the protons more effectively than the electrons. As the gas falls inward and its density ρ rises, the temperature T of each species increases adiabatically as $T \propto \rho^{\gamma-1}$, where γ is the corresponding adiabatic index. At radii $r < 10^2 r_{\rm Sch}$, the electrons are relativistic, with $\gamma = 4/3$, and so their temperature rises inward with increasing density as $T_e \propto \rho^{1/3}$, while the protons are nonrelativistic, with $\gamma = 5/3$, and so $T_p \propto \rho^{2/3}$. The result is a two-temperature plasma in which the protons are much hotter than the electrons. Typical analytic models[10] yield $T_p \sim 10^{12}\ {\rm K}(r/r_{\rm Sch})^{-1}$ and $T_e \sim \min(T_p, 10^{9-11}\ {\rm K})$. Because the typical sound speed is comparable to the Keplerian speed at each radius, the geometry of the flow is thick— making RIAFs the viscous analogs of Bondi accretions.

Analytic models imply a radial velocity that is a factor of $\sim\alpha$ smaller than the free-fall speed and an accretion time that is a factor of $\sim\alpha$ longer than the free-fall time. However, since the sum of the kinetic and thermal energies of a proton is comparable to its gravitational binding energy, RIAFs are expected to be associated with strong outflows.

The radiative efficiency of RIAFs is smaller than the thin-disk value (ϵ). The thin-disk value applies to high accretion rates above some critical value, $\dot{m} > \dot{m}_{\rm crit}$, where $\dot{m}$ is the accretion rate (in Eddington units) near the ISCO, after taking into account the fact that some of the infalling mass at larger radii is lost to outflows. Analytic RIAF models typically admit of a radiative efficiency of

$$\frac{L}{\dot{M}c^2} \approx \epsilon \left(\frac{\dot{m}}{\dot{m}_{\rm crit}}\right) \tag{7.32}$$

for $\dot{m} < \dot{m}_{\rm crit}$, with $\dot{m}_{\rm crit} \sim 0.01$–$0.1$. For example, in the nucleus of the Milky Way, massive stars shed $\sim 10^{-3} M_\odot\ {\rm yr}^{-1}$ of mass into the radius of influence of central black hole Sgr A*, but only a tiny fraction $\sim 10^{-5}$ of this mass accretes onto the black hole.

Since mergers are rare at low redshifts and much of the gas in galaxies has already been consumed in making stars, most local supermassive black holes are characterized by a very low accretion rate. The resulting low luminosity of these dormant black holes, such as the $4 \times 10^6 M_\odot$ black hole lurking at the center of the Milky Way galaxy, is often described using RIAF/ADAF models. Although this mode of accretion is characterized by a low mass infall rate, it could persist over a period of time that is orders of magnitude longer than the quasar mode discussed earlier, so its contribution to the growth of black holes in galactic nuclei may not be negligible.

7.4 The First Black Holes and Quasars

What seeded the formation of supermassive black holes only a billion years after the Big Bang? We know how to make a black hole out of a massive star. When the star ends its life, it stops producing sufficient energy to hold itself against its own gravity, and its core collapses to make a black hole. Long before evidence for black holes was observed, this process leading to their existence was understood theoretically by Robert Oppenheimer and Hartland Snyder in 1937. However, growing a supermassive black hole is more difficult. There is a maximum luminosity at which the environment of a black hole of mass $M_{\rm BH}$ may shine and still accrete gas.[i]

This Eddington luminosity, L_E, was derived in equation (5.27) by balancing the inward force of gravity on each proton by the outward radiation force on its companion electron (which is the momentum flux carried by the radiation times the scattering cross section of the electron) at a distance r:

$$\frac{GM_{\rm BH}m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{7.33}$$

where m_p is the proton mass, and $\sigma_T = 0.67 \times 10^{-24}$ cm^2 is the electron cross section for scattering a photon (Thomson scattering). Interestingly, the limiting luminosity is independent of radius in the Newtonian regime. Since the Eddington luminosity represents an exact balance between gravity and radiation forces, it equals the luminosity of massive stars that are held at rest against gravity by radiation pressure, as described by equation (7.33). This limit is formally valid in a spherical geometry, and exceptions to it were conjectured for other accretion geometries over the years (and we shall consider one such possibility for a spherical geometry in §7.4.1). But, remarkably, observed quasars for which black hole masses can be measured by independent methods appear to respect this limit. When we substitute all constants, the Eddington luminosity is given by

$$L_E = 1.3 \times 10^{44} \left(\frac{M_{\rm BH}}{10^6\ M_\odot}\right) \text{ erg s}^{-1}, \tag{7.34}$$

As discussed previously, we can write the total luminosity from gas accreting onto a black hole, L, as some radiative efficiency ϵ times the mass accretion

[i]Whereas the gravitational force acts mostly on the protons, the radiation force acts primarily on the electrons. These two species are tied together by a global electric field, so that the entire plasma behaves as a single quasi-neutral fluid that is subject to both forces. Under similar circumstances, electrons are confined to the Sun by an electric potential of about a kilovolt (corresponding to a total charge of $\sim$ 75 C). The opposite electric forces per unit volume acting on electrons and ions in the Sun cancel out, so that the total pressure force is exactly balanced by gravity, as for a neutral fluid. An electric potential of 1–10 kV also binds electrons to clusters of galaxies (where the thermal velocities of these electrons, $\sim 0.1c$, are well in excess of the escape speed from the gravitational potential).[11]

rate $\dot{M}$,

$$L = \epsilon \dot{M} c^2, \tag{7.35}$$

where the black hole accretes the nonradiated component, $\dot{M}_{\rm BH} = (1-\epsilon)\dot{M}$. The equation that governs the growth of the black hole mass is then

$$\dot{M}_{\rm BH} = \frac{M_{\rm BH}}{t_E}, \tag{7.36}$$

where (after we substitute all fundamental constants),

$$t_E = 4 \times 10^7 \left(\frac{\epsilon/(1-\epsilon)}{0.1}\right)\left(\frac{L}{L_E}\right)^{-1} \text{ yr}. \tag{7.37}$$

We therefore find that as long as fuel is amply supplied, the black hole mass grows exponentially in time, $M_{\rm BH} \propto \exp\{t/t_E\}$, with an e-folding time t_E. Since the growth time in equation (7.37) is significantly shorter than the $\sim 10^9$ years corresponding to the age of the Universe at a redshift $z \sim 6$—where black holes with a mass $\sim 10^9\ M_\odot$ are found—one might naively conclude that there was plenty of time to grow the observed black hole masses from small seeds. For example, a seed black hole from a Population III star of 100 $M_\odot$ can grow in less than a billion years to $\sim 10^9\ M_\odot$ for $\epsilon \sim 0.1$ and $L \sim L_E$. However, the intervention of various processes makes it unlikely that a stellar mass seed is able to accrete continuously at its Eddington limit without interruption.

For example, mergers were very common in the early Universe. Whenever two gas-rich galaxies come together, their black holes are likely to coalesce. The coalescence is initially triggered by "dynamical friction" from the surrounding gas and stars and is completed—once the binary gets tight—as a result of the emission of gravitational radiation.[12] The existence of gravitational waves is a generic prediction of Einstein's theory of gravity. They represent ripples in space–time generated by the motion of the two black holes as they move around their common center of mass in a tight binary. The energy carried by the waves is taken away from the kinetic energy of the binary, which therefore tightens with time. Computer simulations reveal that when two black holes with unequal masses merge to make a single black hole, the remnant gets a kick owing to the nonisotropic emission of gravitational radiation at the final plunge (see §7.7). This kick was calculated recently using advanced computer codes that solve Einstein's equations (a task that was plagued for decades with numerical instabilities).[13] The typical kick velocity is hundreds of kilometers per second (and up to ten times more for special spin orientations), much larger than the escape speed from the first dwarf galaxies. Thus, continuous accretion was likely punctuated by black hole ejection events[14] that forced the merged dwarf galaxy to grow a new black hole seed from scratch.

7.4.1 Supermassive Stars

If continuous feeding is halted, or if the black hole is temporarily removed from the center of its host galaxy, then one is driven to the conclusion that the black hole seeds must have started more massive than $\sim 100\ M_\odot$. More massive seeds may originate from *supermassive stars*, defined as hydrostatic configurations with masses of 10^3 to $10^8\ M_\odot$. Such systems have not been observed yet. Theoretically, they are expected to be supported almost entirely by radiation pressure and hence their luminosity equals the Eddington limit, $L_E = 1.3 \times 10^{44}(M_\star/10^6\ M_\odot)$ erg s^{-1}. Supermassive stars steadily contract and convert their gravitational binding energy to radiation with a total lifetime $<10^6$ yr before they collapse to a black hole.

Was it possible to make such stars in early galaxies? Yes: numerical simulations indicate that stars of mass $\sim 10^6\ M_\odot$ could have formed at the centers of early dwarf galaxies that were barely able to cool their gas through transitions of atomic hydrogen and had $T_{\rm vir} \sim 10^4$ K and no H_2 molecules (which were suppressed by a Lyman-Werner background). Such systems had a total mass several orders of magnitude higher than the earliest Jeans-mass condensations discussed in §3.2. In both cases, the gas lacks the ability to cool well below $T_{\rm vir}$, and so it fragments into one or two major clumps. The simulation shown in Figure 7.6 results in clumps of several million solar masses, which inevitably form massive black holes. The existence of such massive seeds would have given a jump start to the black hole growth process.

First, we show that the envelope of such stars must be convective. The condition for convective instability is that the star exhibit a negative entropy gradient. This follows from the fact that convective eddies that are hotter and rarefied relative to their environment tend to rise toward the star's surface and decrease their density adiabatically (at constant entropy) in pressure equilibrium with their environment. If the background entropy decreases as the eddies rise, then they become even more rarefied relative to their environment (lower ambient entropy at the same pressure implies higher ambient density) and continue to rise even further, hence leading to an instability. The energy transport by convective eddies drives the star to a state of marginal stability, namely, nearly uniform entropy. Let us first show that in the absence of convection, a supermassive star tends to develop a negative entropy gradient as it radiates away its energy.

The entropy of each electron–proton fluid element in a supermassive star changes according to the local radiative heat flux **F** at a rate

$$\frac{2T_g}{m_p}\frac{\partial s}{\partial t} = -\frac{1}{\rho}\nabla \cdot \mathbf{F}, \tag{7.38}$$

where ρ, T_g, and s are the mass density, temperature, and specific entropy of the element, respectively. If the opacity is dominated by Thomson scattering, then the local radiative heat flux **F** is related to the radiation pressure gradient by

$$\mathbf{F} = \frac{-m_p}{\sigma_T}\frac{1}{\rho}\nabla p_{\rm rad}. \tag{7.39}$$

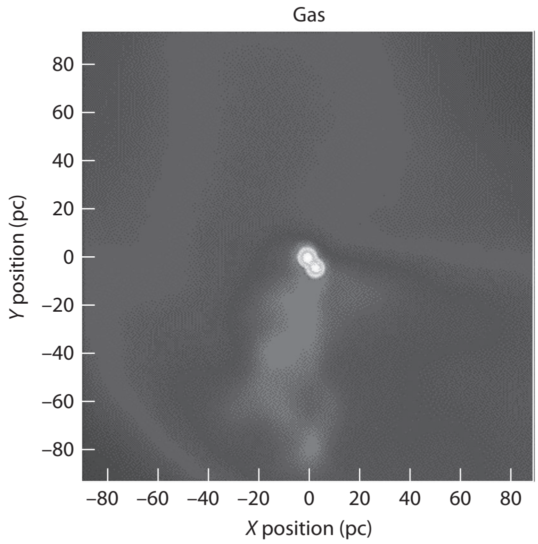

Figure 7.6 Numerical simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2 (see *Color Plate 14* for a color version of this figure). The image shows a snapshot of the gas density distribution 500 million years after the Big Bang, indicating the formation of two compact objects near the center of the galaxy with masses of $2.2 \times 10^6\ M_\odot$ and $3.1 \times 10^6\ M_\odot$, respectively, and radii < 1 pc. Sub-fragmentation into lower mass clumps is inhibited because hydrogen atoms cannot cool the gas significantly below its initial temperature. These circumstances lead to the formation of supermassive stars that inevitably collapse to make massive seeds for supermassive black holes. The simulated box size is 200 pc on a side. Bromm, V., & Loeb, A., *Astrophys. J.* **596**, 34 (2003). Reproduced by permission of the AAS.

If we ignore gas pressure and rotation, the hydrostatic equilibrium equation is simply

$$\frac{1}{\rho}\nabla p_{\rm rad} = \mathbf{g}, \tag{7.40}$$

where the gravitational field $\mathbf{g}$ obeys Poisson's equation,

$$\nabla \cdot \mathbf{g} = -4\pi G\rho. \tag{7.41}$$

Combining equations (7.39)–(7.41), we find that the right-hand side of equation (7.38) is constant,

$$\frac{1}{\rho}\nabla \cdot \mathbf{F} = \frac{4\pi G m_p}{\sigma_T} = \text{constant}. \tag{7.42}$$

Therefore, the gradient of equation (7.38) gives

$$\frac{\partial}{\partial t}\nabla s = \frac{2\pi G m_p^2}{\sigma_T}\frac{\nabla T_g}{T_g^2} < 0. \quad (7.43)$$

The radial temperature gradient is negative, since heat flows out of the star, which implies that the star will develop a negative entropy gradient and become convectively unstable. This result holds also for a rotating star, as long as the rotation period is much longer than the dynamical time, $t_{\rm dyn} \sim (G\rho)^{-1/2} = 1.1\ {\rm hr}(\rho/1\ {\rm g\ cm^{-3}})^{-1/2}$.

The nearly uniform entropy established by convection makes the structure of supermassive stars simple (equivalent to a so-called polytrope with an index $n = 3$) with a unique relation between their central temperature T_c and central density ρ_c,[15]

$$T_c = 2\times 10^6\ {\rm K}\left(\frac{\rho_c}{1\ {\rm g\ cm^{-3}}}\right)^{1/3}\left(\frac{M_\star}{10^6\ M_\odot}\right)^{1/6}. \quad (7.44)$$

Because of this modest temperature, nuclear reactions are insignificant in metal-poor stars with masses $M_\star > 10^5\ M_\odot$. General-relativistic corrections make the star unstable to direct collapse to a black hole as soon as its radius contracts to a value

$$R_\star < R_{\rm crit} = 1.59\times 10^3\left(\frac{M_\star}{10^6\ M_\odot}\right)^{1/2}\left(\frac{GM_\star}{c^2}\right). \quad (7.45)$$

Rotation can stabilize supermassive stars to smaller radii, but even rotating stars are expected to eventually collapse to a black hole after shedding their angular momentum through a wind. If the supermassive star is made of preenriched gas, then powerful winds will inevitably be driven at its surface, where the opacity due to lines from heavy elements far exceeds the Thomson value, making the outward radiation force stronger than gravity (see §7.5).

We note that the infall of a sufficiently dense, optically thick spherical envelope of gas cannot be prevented by radiation pressure even if the radiation production rate exceeds the Eddington limit near the center. To see this, let us consider a gas shell falling inward with a velocity $v_{\rm in}(r)$ at a radius r. If the outward diffusion time of photons through the gas, $t_{\rm diff} \sim \tau r/c$ (where $\tau \sim \sigma_T \rho r/m_p$ is the shell's optical depth to Thomson scattering), exceeds the infall time, $t_{\rm in} \sim r/v_{\rm in}$, then the radiation will be dragged by the infalling gas into the black hole. Even though the radiation is diffusing outward in the local rest frame of the gas, it is moving inward in the black hole frame of reference when $t_{\rm diff} > t_{\rm in}$. In that regime, the radiation will never be able to counteract the collapse of gas shells that are farther out. Expressing the mass accretion rate as

$$\dot{M} = 4\pi\rho r^2 v_{\rm in}, \quad (7.46)$$

we find that $t_{\rm diff} > t_{\rm in}$ if

$$\frac{\dot{M}}{\dot{M}_{\rm E}} > \epsilon\left(\frac{r}{GM/c^2}\right), \quad (7.47)$$

where M is the mass interior to radius r, and $\dot{M}_{\rm E} = L_{\rm E}/\epsilon c^2$ is the mass accretion rate that produces the Eddington luminosity (equation 7.34) for a radiative efficiency ϵ. We therefore conclude that as long as the mass infall rate is sufficiently high, the Eddington limit will not apply because of photon trapping. Super-Eddington accretion can therefore grow a seed black hole rapidly, as long as the blanket of infalling gas advects the radiation inward as it accretes onto the black hole. This "obscured" mode of black hole accretion (which is hidden from the view of observers) could be particularly important at high redshifts when the gas density and infall rate onto galaxies ottain their highest values.

7.5 Black Holes and Galaxies

7.5.1 The Observed Correlations between Supermassive Black Holes and Host Galaxies

As described in §7.1, there are clear correlations between the supermassive black holes that reside at the centers of galaxies and their host spheroids of stars. In particular, Figure 7.7 shows the strong correlation between the mass of the central black hole and the velocity dispersion of stars in the host galaxy's spheroid, $\sigma_\star$. (Note that this is *not* necessarily identical with the virial velocity of the galaxy, because it samples motions on much smaller scales.) The data at $z \sim 0$ follow a tight correlation, $M_{\rm BH} \propto \sigma_\star^4$, with only a small apparent intrinsic scatter. Other correlations also exist; especially important is the observed relationship between the black hole mass and spheroid luminosity at $z \sim 0$, $M_{\rm BH} \propto L_{\rm sp}$, and that between the black hole mass and spheroid mass, $M_{\rm BH} \sim 0.0014 M_{\rm sp}$.[16]

These relationships explain how quasars may shine much brighter than their host galaxies. A typical star like the Sun emits a luminosity $L_\odot = 4 \times 10^{33}$ erg s^{-1}, which can also be written as a fraction $\sim 3 \times 10^{-5}$ of its Eddington luminosity $L_E = 1.3 \times 10^{38}$ erg s^{-1}. At the present day, black holes grow up to a fraction $\sim 10^{-3}$ of the stellar mass of their spheroid. When they shine close to their Eddington limit, they may therefore outshine their host galaxy by up to a factor of $\sim 10^{-3}/(3 \times 10^{-5})$, namely, one to two orders of magnitude. The factor is smaller during short starburst episodes, which are dominated by massive stars with larger Eddington fractions.

Measuring black hole masses and host galaxy properties at high redshift is difficult, so the extrapolation of these relationships toward the cosmic dawn is not yet clear. The best way to do so is to resolve directly the region of influence of the black hole, where its gravitational potential dominates the motion of stars. Crudely, its mass can then be measured by fitting a Keplerian rise to the velocity dispersion at decreasing separations from the center. In practice, this is difficult because of the complex structure of galaxies and the projection of stars outside the region into the measured region. Nevertheless, the technique has been used successfully in dozens of nearby galaxies, as shown in Figure 7.7.

At higher redshifts, such direct observations are impossible, so indirect techniques are necessary. These rely on local "virial relations" between black hole

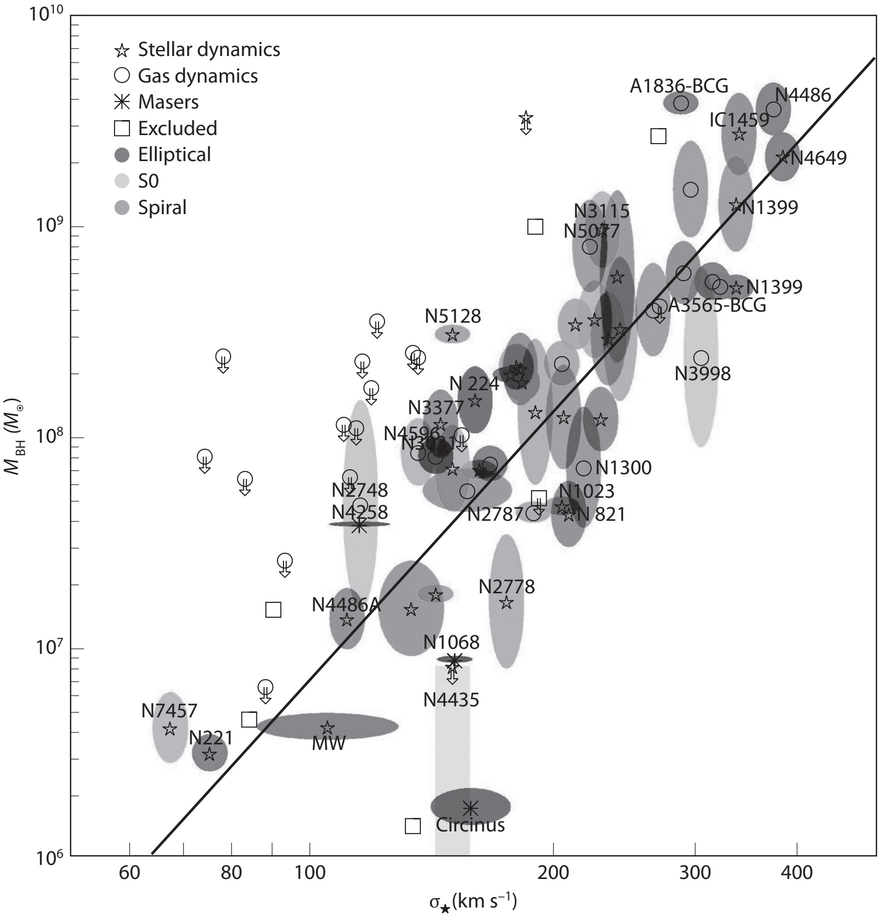

Figure 7.7 Dynamical measurements of the correlation between supermassive black hole mass, $M_{\rm BH}$, and velocity dispersion of stars in the spheroid of its host galaxy, $\sigma_\star$ (see *Color Plate 15* for a color version of this figure). The symbol indicates the method of black hole mass measurement: dynamics of stars (*pentagrams*), dynamics of gas (*circles*), dynamics of maser sites (*asterisks*). Arrows indicate 3σ upper limits to black hole mass. The shade of the error ellipse indicates the Hubble type of the host galaxy: elliptical, S0, or spiral. The line is the best-fit relation to the full sample: $M_{\rm BH} = 10^{8.12}\ M_\odot(\sigma_\star/200\ {\rm km\ s^{-1}})^{4.24}$. The mass uncertainty for NGC 4258 has been plotted much larger than its actual value to show on this plot. Gültekin, K., et al., *Astrophys. J.* **698**, 198 (2009). Reproduced by permission of the AAS.

mass and spectroscopic properties of the AGN. These luminous sources typically show broad emission lines that (based on atomic physics considerations) are believed to originate from dense, highly ionized gas clouds near the black hole. Assuming that the velocity dispersion of these clouds v (which is of the order of a few thousand kilometers per second) is gravitationally induced by the central black hole at their characteristic orbital radius r, we have $v^2 \sim GM_{\rm BH}/r$. The velocity can be measured from the Doppler width of the corresponding emission line. However, to obtain $M_{\rm BH}$ one also needs a measurement of the orbital radius of these clouds, which can be obtained through *reverberation mapping* observations that monitor the AGN spectrum over time. If the continuum emission (from the inner region of the accretion disk around the black hole) increases, then the broad-line region emission should also increase—but there will be a time delay as the continuum light propagates to that region, which can be measured with intensive monitoring programs. Cosmic time dilation makes this measurement more difficult for massive black holes (which are the only ones that can be easily observed) at high redshift.

Instead, an indirect proxy for the orbital radius is usually used. Empirically, the broad-line regions of all AGN are characterized by roughly the same UV flux, $L/4\pi r^2 \sim$ const. A possible interpretation of this scaling is that the emission lines originate near the outermost boundary where dust sublimates (and the cloud gas is exposed to the central UV illumination), which determines a particular threshold for the UV flux.[17] Thus, $r \propto L^{1/2}$ (with a proportionality constant that is fixed by the geometry and kinematics of the region), and the radius r can simply be estimated from the AGN luminosity. Measurements of nearby AGN are used to calibrate the proportionality constant.[18] The local AGN samples focus on emission lines easily accessible to optical telescopes observing at $z \sim 0$ (such as Hβ); extension to higher redshifts requires rest-UV lines (such as C IV) that redshift into the optical range. The calibration is not trivial, but it provides at least an indirect route to estimating black hole masses at high redshifts.

These techniques allow an estimate of the black hole mass, but to understand the relationship between the black hole and its host galaxy, one must also measure that galaxy's properties. This is also very difficult, because the only quasars visible at high redshifts are themselves extremely luminous—and hence dramatically outshine their host galaxies. Extremely good angular resolution is required to distinguish the pointlike AGN from the surrounding stars and gas (which is easiest to achieve for quasars that are magnified by intervening gravitational lenses). This measurement is most feasible in the radio, where the molecular line widths also provide a reasonably good estimate for the velocity dispersion and mass of the central star-forming portion of the galaxy. Assuming that we can classify that region as a spheroid, the highest-redshift quasars appear to have much more massive black holes than the local relations would suggest, with $M_{\rm BH} \sim 0.02 M_{\rm sp}$. The data are so far sparse, but interestingly, there appears to be a much weaker correlation with $\sigma_\star$ than in the local Universe.[19]

7.5.2 Galaxies and Their Supermassive Black Holes

Regardless of their extension to higher redshifts, the tight correlations in the local Universe described in §7.5.1 suggest an intimate connection between the growth of galaxy spheroids (or bulges) and their central black holes. This relation in turn suggests that either feedback from the black hole affects the galaxy's gas and stars, or feedback from those stars affects the black holes—or both. Both cases suggest that the relationship is an extremely important one for galaxy formation and evolution.

Since the mass of a galaxy at a given redshift scales with its virial velocity as $M_g \propto V_c^3$ (see equation 3.31), the binding energy of galactic gas scales as $M_g V_c^2 \propto V_c^5$. Meanwhile, the momentum required to expel the gas from its host scales as $M_g V_c \propto V_c^4$. Both scalings are reasonably close to the observed correlation shown in Figure 7.7 and can be tuned to explain the observed relations by appealing to feedback from the black hole to shut off star formation in the galaxy.

More concretely, suppose that accretion onto a black hole launches a wind that couples to the galaxy's ISM, driving it out of the potential well. As an example, suppose that this feedback contains a total energy that is a fraction ϵ_w of the rest mass of the black hole. Using similar arguments to those leading to equation (6.26), we find that this energy can unbind the gas in the galaxy once

$$M_{\rm BH} \sim 2 \times 10^4 \left(\frac{0.05}{\lambda} \frac{0.1}{\epsilon_w} \right) \left(\frac{M_h}{10^{10}\ M_\odot} \right)^{5/3} \left(\frac{1+z}{10} \right) M_\odot \tag{7.48}$$

$$\sim 2 \times 10^4 \left(\frac{0.05}{\lambda} \frac{0.1}{\epsilon_w} \right) \left(\frac{V_c}{78 \text{ km s}^{-1}} \right)^{5} \left(\frac{1+z}{10} \right)^{-3/2} M_\odot, \tag{7.49}$$

where we have written the limit in terms of both the halo mass and circular velocity for convenience. One can then imagine that because this same gas is ultimately the fuel source for the black hole, both it and the galaxy's stellar population will stop growing once it reaches this limiting size, establishing the tight underlying correlation. Importantly, this limiting size implies that black holes—like supernovae—are much more efficient at expelling gas from small halos than large ones, so black holes may also play a very significant role at high redshifts.

Similarly, following the arguments leading to equation (6.31), we see that the momentum input from a black hole wind will suffice to drive the gas out of a galaxy if[ii]

$$M_{\rm BH} \sim 6 \times 10^6 \left(\frac{0.1}{\epsilon_w} \right) \left(\frac{M_h}{10^{10}\ M_\odot} \right)^{4/3} \left(\frac{1+z}{10} \right) M_\odot \tag{7.50}$$

$$\sim 6 \times 10^6 \left(\frac{0.1}{\epsilon_w} \right) \left(\frac{V_c}{78 \text{ km s}^{-1}} \right)^{4} \left(\frac{1+z}{10} \right)^{-1} M_\odot. \tag{7.51}$$

As expected, this approach provides a more stringent limit because it assumes that the outflow loses energy as it propagates through the galaxy.

This general picture is an attractive one, but note that neither of these simple scenarios has a straightforward connection to the observed correlations. This

[ii] Here we use $V_c^2 = 2\sigma_\star^2$ for a singular isothermal sphere profile, and we assume that the black hole releases most of its energy over a dynamical time $\sim \sigma_{\rm vir}/r_\star$.

is easiest to see by noting that although the scaling with velocity could easily be *interpreted* as mimicking the observed scaling with $\sigma_\star$, that would require a nonlinear relationship between spheroid mass (or luminosity) and halo mass. A key question for understanding how black holes interact with their hosts is then obviously how the spheroid properties like $\sigma_\star$ relate to the halo properties. One simple correction may be to note that if the fuel supplies of stellar bulges and supermassive black holes are identical, a linear relation between bulge mass and black hole mass may be naturally described by the feedback-regulated growth picture described. In that case, the relation between bulge mass and velocity dispersion is an interesting one that reflects how galaxies accumulate and process cold gas.

This relationship then raises an interesting "chicken and egg" problem, because *both* the black hole and stellar component generate feedback on the gas supply, and it is difficult to diagnose from observations which is more significant. The energy injected by supernovae of bulge stars is $\sim \omega_{\rm SN} M_{\rm sp}$, where $M_{\rm sp}$ is the stellar mass of the bulge component, and $\omega_{\rm SN}$ is the energy output per unit mass in stars, $\sim 10^{49}$ erg $M_\odot^{-1}$ for typical stellar populations. The energy injected by the black hole is $\sim \epsilon_w f_{\rm BH} M_{\rm sp} c^2$, where $f_{\rm BH}$ is the fraction of the stellar bulge mass in the black hole ($\sim$0.0014 in the nearby Universe, and possibly increases in massive halos at higher redshift; see §7.5.1). Then, the energy output from the black hole dominates if

$$\epsilon_w > 0.004 \left(\frac{0.0014}{f_{\rm BH}} \right) \left(\frac{\omega_{\rm SN}}{10^{49} \text{ erg } M_\odot^{-1}} \right), \tag{7.52}$$

which is clearly very small. However, as we discuss in §7.5.3, it is not clear how this energy reservoir couples to the galaxy ISM.

The answers to these questions depend on the way in which black holes (and the central stellar regions) are fed by gas in their host galaxies, which requires either a way to dissipate the angular momentum of the gas or an external force that can torque the gas toward the center. Viscous dissipation (even from instabilities like the MRI described in §7.3.2) is quite slow in the ISM and is not thought sufficient to feed an AGN.

Instead, the most popular model is one in which supermassive black holes in galaxies are fed with gas in episodic events of gas accretion triggered by mergers of galaxies. Tides and gravitational interactions during major mergers (i.e., in which the two merging components have comparable mass) provide massive torques that efficiently drive a large fraction of the gas toward the center of the merging components. However, black hole growth does not *require* a merger; any other process that can drive gas toward the galaxy center will also suffice, such as global disk instabilities. These can themselves be triggered by accretion streams that feed the galaxy halo itself at a high rate.

In the merger case, the fundamental cosmological input to modeling the growth of black holes is the merger rate of dark matter halos. For example, it is possible to "dress up" the mass distribution of halos in Figure 3.10 using the excursion set model to estimate their merger rates (see §3.4.2). The key physical inputs are the black hole–halo mass relation and a light curve for the quasar.

The former can be fixed via the observed $M_{\rm BH}$–$\sigma_\star$ relation. The latter then determines the (observable) luminosity from the black hole properties. In its most crude form, this value requires a maximum luminosity (which can be set as a fixed fraction of the Eddington limit, for example) and a duty cycle (which may be related to the black hole growth time t_E of equation 7.37 or to the dynamical time of the host galaxy, which is approximately the merger timescale). This simple approach can be tuned to give good agreement with the data on the quasar luminosity function shown in Figure 7.1. In the crudest form of the model, reasonable agreement with observations requires that quasars deposit $\sim$5% of their Eddington luminosity in the ISM of their host galaxy.[20]

Regardless of its origin, the inflow of cold gas toward galaxy centers during the growth phase of their black holes is naturally accompanied by a burst of star formation. The fraction of gas not consumed by stars or ejected by supernova-driven winds continues to feed the black hole. It is therefore not surprising that quasars and starbursts coexist in ultraluminous galaxies, and that all quasars show strong spectral lines of heavy elements.[21] In fact, as discussed in §7.3.2, the outskirts of accretion disks are very likely gravitationally unstable and susceptible to fragmentation to form stars.

7.5.3 Jets and Winds: Black Hole Feedback Mechanisms

Although the general picture of black holes' exerting strong feedback on their host galaxies is very attractive, the details remain vague, primarily because little direct evidence exists that any given feedback mechanism affects the host galaxy. The coupling mechanism is unknown, though several plausible candidates do exist: the process could be related to either the bright radiation or fast outflows that are known to be produced by quasars.

One possibility traces back to the fact that the scattering cross section per unit mass for UV radiation on dust is larger than σ_T/m_p (which is used to set the Eddington luminosity in equation 7.34) by two orders of magnitude.[22] Although dust is destroyed within $\sim 10^4 GM_{\rm BH}/c^2$ by the strong illumination from an Eddington-limited quasar, it should survive at larger distances. Hence, the radiation pressure on dust would exceed the gravitational force toward the black hole and drive powerful outflows.

Spectral lines could be even more effective than dust in their coupling to radiation. The integral of the absorption crosssection of a spectral line over frequency,

$$\int \sigma(\nu)\, d\nu = f_{\rm osc}\left(\frac{\pi e^2}{m_e c}\right), \tag{7.53}$$

where $f_{\rm osc}$ is the absorption oscillator strength, is typically orders of magnitude larger than the Thomson value. For example, the Lyman-α transition of hydrogen, for which $f_{\rm osc} = 0.416$ (see §4.2), provides an average cross section that is seven orders of magnitude larger than σ_T when averaged over a frequency band as wide as the resonant frequency itself. Therefore, lines could be even more effective at driving outflows in the outer parts of quasar environments.

A second possibility for launching outflows is a (magneto)hydrodynamical effect within the accretion disk. We have already described how radiatively inefficient accretion flows can launch outflows, but there is good evidence for outflows in strongly emitting AGN as well. These outflows take (at least) two types. The first are jets, as shown in Figure 7.2, which are thought to be launched from the inner regions of accretion disks owing to complex magnetohydrodynamical effects. The material inside travels relativistically and may carry a great deal of energy, comparable to the radiative luminosities of the sources. However, such jets have extremely narrow opening angles and do not appear to interact with much of the galaxy's ISM. They are therefore unlikely to provide efficient feedback. The second possibility is a more classic "wind" launched from the disk itself, perhaps along open magnetic field lines threading the disk.

However they are launched, there is good evidence for these winds in some quasars. Called *broad absorption line quasars* (BALQSOs), these objects have strong redshifted absorption lines with extremely large widths $v_{\rm BAL} \sim 0.1c$ that are thought to represent outflowing material. The total kinetic luminosity in these winds is difficult to measure:

$$L_K \approx 2\pi f_c N_{\rm H} m_p v_{\rm BAL}^3 R_{\rm BAL}, \tag{7.54}$$

where f_c is the covering factor of the outflow, $N_{\rm H}$ is the column density of material in the absorber, and $R_{\rm BAL}$ is its distance from the central source. The covering fraction is uncertain, but a "unified" quasar model in which BALQSOs differ from "normal" quasars only because of the observer's viewing angle implies that $f_c \sim 0.1$. The radius of the absorbing material is even more uncertain, relying on modeling of the gas or outflow shock. Models suggest that $\epsilon_w \sim 0.03$ for BALQSOs, though with a large uncertainty.[23]

Although a wind can clearly couple to the ISM better than a narrow jet, it is still not obvious how the wind propagates from the central AGN point source through the much larger star-forming region. A great deal of work is required to understand the physics of quasar outflows and their influence on their host galaxies.

7.5.4 Black Holes and "External" Feedback

So far, we have focused on the interplay of black holes and their host galaxies. However, the possibility that black holes can expel large fractions of the gas from their hosts also indicates that their feedback mechanisms may influence the surrounding IGM and halos. Energetically, equation (7.52) suggests that even a modest coupling between feedback energy and the ISM or IGM suffices to make AGN outflows important relative to the radiative or supernova feedback from their host. There is even some direct evidence for AGN feedback in the local Universe: many galaxy clusters contain hot bubbles of material fed by radio jets from their massive central galaxies. Clusters also provide the strongest indirect evidence for intergalactic AGN feedback, where AGN are often invoked to explain the lack of apparent cooling in the cluster gas. (More generally, models often appeal to them to avoid the "overcooling problem" described in §8.2, though there is little direct evidence for their activity in dark matter halos smaller than groups of galaxies.)

Fortunately, these large-scale black hole outflows can be modeled in the same way as their stellar counterparts (see §6.5.1), because of the approximate self-similarity of blast waves in an expanding Universe. AGN outflows will have many of the same effects as stellar superwinds, though the details will, of course, differ. For example, quasar winds may not carry as much of the host ISM out of the galaxy as their stellar counterparts and will therefore cause less IGM metal enrichment. However, plasmas launched from the highly magnetized accretion disk may drag magnetic fields with them, seeding an intergalactic magnetic field at high redshifts.

7.6 Black Hole Binaries

Nearly all nearby galactic spheroids are observed to host a nuclear black hole. Therefore, the hierarchical buildup of galaxies through mergers must generically produce black hole binaries. Such binaries tighten through dynamical friction with the background gas and stars and ultimately coalesce through the emission of gravitational radiation.

When a tight binary is created from a merger of separate galaxies, the mass ratio of two black holes cannot be too extreme. A satellite of mass $M_{\rm sat}$ in a circular orbit at the virial radius of a halo of mass $M_{\rm halo}$ will sink to the center in a dynamical friction time of $\sim 0.1 t_H (M_{\rm halo}/M_{\rm sat})$, where t_H is the Hubble time. If the orbit is eccentric with an angular momentum that is a fraction ε of a circular orbit with the same energy, then the sinking time is reduced by a factor of[24] $\sim\varepsilon^{0.4}$. Therefore, massive satellites with $M_{\rm sat} > 0.1 M_{\rm halo}$ bring their supermassive black holes to the center of their host halos over the age of the Universe.

As a satellite galaxy sinks, its outer envelope of dark matter and stars is stripped by tidal forces. The stripping is effective down to a radius inside which the mean mass density of the satellite is comparable to the ambient density of the host galaxy.[25] Eventually, the two black holes are stripped down to the cores of their original galaxies and are surrounded by a circumbinary envelope of stars and gas. As long as the binary is not too tight, the reservoir of stars within the binary orbit can absorb the orbital binding energy of the binary and allow it to shrink. However, when the orbital velocity starts to exceed the local velocity dispersion of stars, a star impinging on the binary is typically expelled from the galactic nucleus at a high speed. This expulsion occurs at the so-called hardening radius of the binary,

$$a_{\rm hard} \approx 0.1 \frac{q}{(1+q)^2} M_6 \left(\frac{\sigma_\star}{100\ {\rm km\ s^{-1}}}\right)^{-2}\ {\rm pc}, \tag{7.55}$$

at which the binding energy per unit mass of the binary exceeds $(3/2)\sigma_\star^2$, where $\sigma_\star$ is the velocity dispersion of the stars before the binary tightened. Here, $M \equiv (M_1 + M_2)$ is the total mass, M_1 and M_2 are the masses of the two black holes,

$M_6 = (M/10^6 M_\odot)$, $q = M_1/M_2$ is the mass ratio. We also let $\mu = M_1 M_2/(M_1 + M_2)$ be the reduced mass of the binary.

A hard binary will continue to tighten only by expelling stars that cross its orbit. Unless the lost stars are replenished by new stars scattered into an orbit that crosses the binary (through dynamical relaxation processes in the surrounding galaxy, whose relaxation time is typically very long), the binary will stall. This "final parsec problem" is circumvented if gas streams into the binary from a circumbinary disk. Indeed, the tidal torques generated during a merger extract angular momentum from any associated cold gas and concentrate the gas near the center of the merger remnant, where its accretion often results in a bright quasar.

If the two black holes are in a circular orbit of radius $a < a_{\rm hard}$ around each other, their respective distances from the center of mass are $a_i = (\mu/M_i)a$, where $i = 1, 2$. We define the parameter $\zeta = 4\mu/(M_1 + M_2)$, which equals unity if $M_1 = M_2$ and is smaller otherwise. The orbital period is given by

$$P = 2\pi(GM/a^3)^{-1/2} = 1.72 \times 10^{-2}\, a_{14}^{3/2} M_6^{-1/2}\ \text{yr}, \tag{7.56}$$

where $a_{14} \equiv (a/10^{14}\ \text{cm})$. The angular momentum of the binary can be expressed in terms of the absolute values of the velocities of its members v_1 and v_2 as $J = \Sigma_{i=1,2} M_i v_i a_i = \mu v a$, where the relative orbital speed is

$$v = v_1 + v_2 = (2\pi a/P) = 1.15 \times 10^4 M_6^{1/2} a_{14}^{-1/2}\ \text{km s}^{-1}. \tag{7.57}$$

In gas-rich mergers, the rate of inspiral decreases as soon as the gas mass interior to the binary orbit falls below μ, when the enclosed gas mass is no longer sufficient to carry away the entire orbital angular momentum of the binary, J. Subsequently, momentum conservation requires that fresh gas steadily flow toward the binary orbit for it to shrink. The binary tightens by expelling gas out of a region twice as large as its orbit (similarly to a blender's opening a hollow gap) and by torquing the surrounding disk through spiral arms. Fresh gas reenters the region of the binary as a result of turbulent transport of angular momentum in the surrounding disk. Since the expelled gas carries a specific angular momentum $\sim va$, the coalescence time of the binary is inversely proportional to the supply rate of fresh gas into the binary region. In a steady state, the mass supply rate of gas that extracts angular momentum from the binary, $\dot{M}$, is proportional to the accretion rate of the surrounding gas disk. Given that a fraction of the mass that enters the central gap accretes onto the black holes and fuels quasar activity, it is appropriate to express $\dot{M}$ in Eddington units, $\dot{m} \equiv \dot{M}/\dot{M}_E$, corresponding to the accretion rate required to power the limiting Eddington luminosity with a radiative efficiency of 10%, $\dot{M}_E = 0.023\, M_\odot\ \text{yr}^{-1} M_6$. We then find that the gas-driven coalescence timescale is

$$t_{\rm gas} \approx (J/\dot{M} v a) = \mu/\dot{M} = 1.1 \times 10^7\, \zeta \dot{m}^{-1}\ \text{yr}. \tag{7.58}$$

For a steady $\dot{m}$, the binary spends equal amounts of time inside each logarithmic interval of radius a until gravitational waves start to dominate its loss of angular momentum.

The coalescence timescale due to gravitational wave emission is given by[26]

$$t_{\rm GW} = \frac{5}{256}\frac{c^5 a^4}{G^3 M^2 \mu} = 2.53 \times 10^3 \,\frac{a_{14}^4}{\zeta M_6^3}\ \text{yr}. \tag{7.59}$$

By setting $t_{\rm GW} = t_{\rm gas}$, we can solve for the orbital speed, period, and separation at which gravitational waves take over:

$$v_{\rm GW} = 4.05 \times 10^3\, \zeta^{-1/4} (\dot{m} M_6)^{1/8}\ \text{km s}^{-1}; \tag{7.60}$$

$$P_{\rm GW} = 0.4\, \zeta^{3/4} M_6^{5/8} \dot{m}^{-3/8}\ \text{yr}; \tag{7.61}$$

$$a_{\rm GW} = 2.6 \times 10^{-4}\, \zeta^{1/2} M_6^{3/4} \dot{m}^{-1/4}\ \text{pc}. \tag{7.62}$$

For a binary redshift z, the observed period is $(1+z)P_{\rm GW}$. The orbital speed at which gravitational waves take over is very weakly dependent on the supply rate of gas, $v_{\rm GW} \propto \dot{M}^{1/8}$, and generically corresponds to an orbital separation of the order of $\sim 10^3\, r_{\rm Sch}$. The probability of finding binaries deeper in the gravitational wave–dominated regime, $\mathcal{P} \propto t_{\rm GW}$, diminishes rapidly at increasing orbital speeds, with $\mathcal{P} \propto (v/v_{\rm GW})^{-8}$.

Black hole binaries can be identified visually or spectroscopically. At large separations the cores of the merging galaxies can easily be identified as separate entities. If both black holes are active simultaneously, then the angular separation between the brightness centroids can in principle be resolved at X-ray, optical, IR, or radio wavelengths. The UV illumination by a quasar usually produces narrow lines from gas clouds at kiloparsec distances within its host galaxy or broad lines from denser gas clouds at subparsec distances from it. Therefore, the existence of a binary can be inferred from various spectroscopic offsets: *(i)* between two sets of narrow lines if the galaxies are separated by more than a few kiloparsecs and both have quasar activity at the same time; *(ii)* between the narrow emission lines of the gas and the absorption lines of the stars due to the tidal interaction between the galaxies at a multikiloparsec separation; *(iii)* between narrow lines and broad lines if the black hole binary separation is between the kilopersec and parsec scales.

The last of these offset signatures also can be produced by a single quasar that gets kicked out of the center of its host galaxy while carrying the broad-line region with it. Such a kick can be produced either by the anisotropic emission of gravitational waves during the coalescence of a binary (which generates a recoil of up to $\sim$200 km s^{-1} in a merger of nonspinning black holes, and up to $\sim$4,000 km s^{-1} for special spin orientations), or from triple black hole systems that form when a third black hole is added to a galaxy center before the binary there has coalesced.[27] Aside from testing general relativity in the strong field limit, fast recoils have an important feedback effect in forcing a fresh start for the growth of black holes in small galaxies at high redshifts. The expected kick velocities of these early recoils offer a very interesting possibility for fossil signatures in the local Universe, because they are much larger than the escape speeds of the dwarf galaxy hosts but comparable to or smaller than the escape speeds of these galaxies' descendants, like the Milky Way. For example, the hierarchical formation of the Milky Way may have left recoiled black holes floating in its halo, which are detectable through the compact star clusters that remain bound to these intermediate-mass black holes following their ejection from their host dwarf galaxies at high redshifts.[28]

7.7 Gravitational Waves from Black Hole Mergers

As just described, the final phase of black hole binary coalescence is driven by the emission of gravitational waves. The emitted waves could be detected by new observatories that are currently being planned or constructed.

As long as the binary separation is much larger than its Schwarzschild radius, the emitted gravitational wave luminosity can be derived in the post-Newtonian approximation. For two black holes on a circular orbit, the luminosity is[29]

$$L_{\rm GW} = \frac{32}{5}\frac{G^4}{c^5}\frac{M^3\mu^2}{a^5}. \tag{7.63}$$

The loss of energy to the emitted waves leads to a decrease in the binary separation a and an eventual coalescence of the two black holes after a time $t_{\rm GW}$ given by equation (7.59). Supermassive binaries with comparable mass members merge in less than a Hubble time once their separation shrinks to $a < 10^{3.5} r_{\rm Sch}$ (see equation 7.2) or once their relative orbital velocity $v = (GM/a)^{1/2} > 10^{-2}c = 3 \times 10^3$ km s^{-1}.

Future detectors will be sensitive to the gravitational wave amplitude. To an order of magnitude, the observed wave amplitude from an equal-mass binary with a Schwarzschild radius $r_{\rm Sch}$ and an orbital velocity v is given by $h \sim (1+z)(r_{\rm Sch}/d_L)(v^2/c^2)$, where d_L is the luminosity distance to the binary.[30] Since the signal amplitude declines only as the inverse distance rather than the inverse distance squared (as for electromagnetic detectors that respond to photon flux), the first generation of sensitive gravitational wave observatories will be able to find sources at cosmological distances.

More accurately, in a reference frame centered on the solar system's barycenter, the gravitational wave amplitude in its two polarization states is given by[31]

$$h_+ = \frac{2\mathcal{M}_z^{5/3}[\pi f_{\rm obs}]^{2/3}}{d_L}\left[1 + (\hat{\mathbf{L}}\cdot\hat{\mathbf{n}})^2\right]\cos[2\Phi(t)]; \tag{7.64}$$

$$h_\times = -\frac{4\mathcal{M}_z^{5/3}[\pi f_{\rm obs}]^{2/3}(\hat{\mathbf{L}}\cdot\hat{\mathbf{n}})}{d_L}\sin[2\Phi(t)]; \tag{7.65}$$

where the so-called redshifted chirp mass $\mathcal{M}_z \equiv (1+z)\mu^{3/5}/M^{2/5}$ sets the rate at which the binary shrinks, determining the "chirp" of its orbital frequency $P = 2\pi/\sqrt{GM/a^3}$. The precise orbital phase of the binary $\Phi(t)$ then depends on the masses and spins of the binary members and yields the observed wave frequency $f_{\rm obs}(t) = [\pi]^{-1}(d\Phi/dt)$, which is $(1+z)$ times smaller than the emitted wave frequency. The unit vector $\hat{\mathbf{n}}$ points from the solar system frame to the binary—defining the sky coordinates of the source, and the unit vector $\hat{\mathbf{L}}$

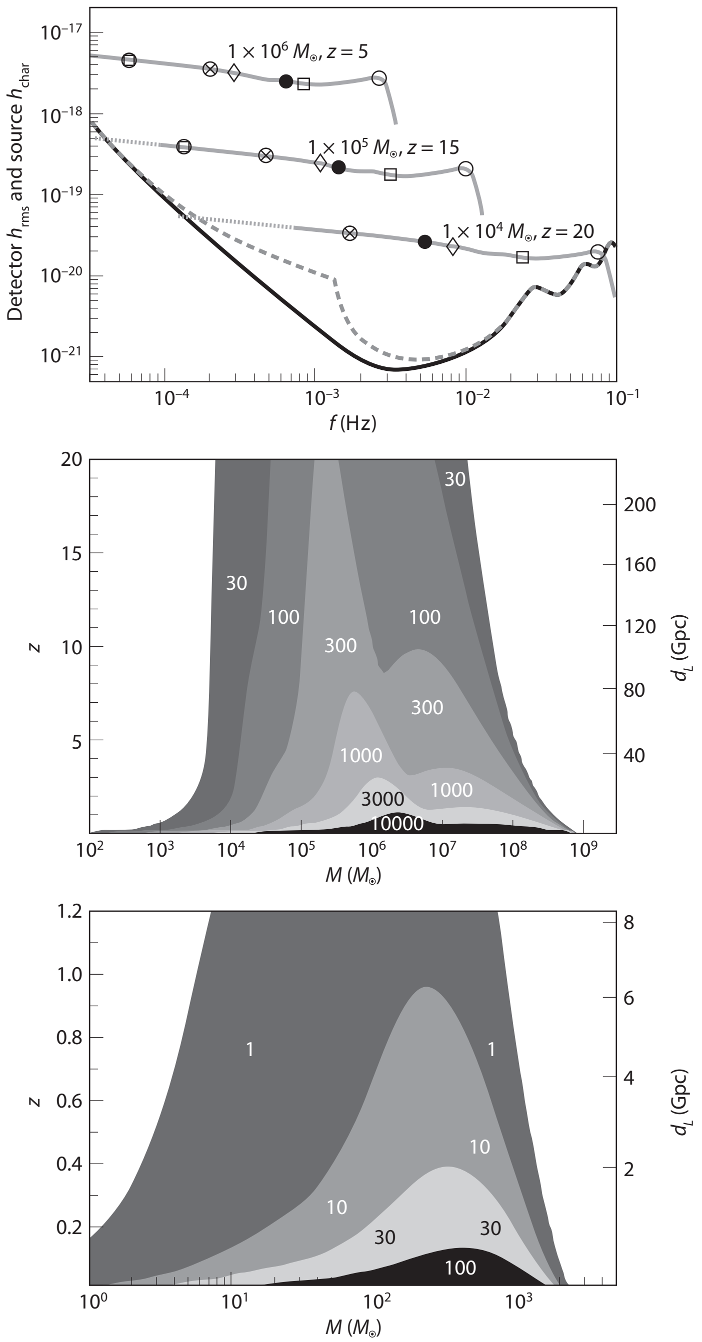

Figure 7.8 Sensitivity of the future gravitational wave observatories, LISA and Advanced-LIGO, to equal-mass ($M_1 = M_2 = M/2$) binaries (see *Color Plate 16* for a color version of this figure). *Top panel:* Root-mean-square noise

points along the direction of the binary angular momentum vector—defining the binary orientation relative to the line of sight. The inspiral signal does not explicitly provide the cosmological redshift separately from the binary masses, but the redshift can be inferred from $d_L(z)$ (or from an electromagnetic counterpart to the gravitational wave signal). Any particular detector is sensitive to a linear combination of the two polarization signals, with coefficients that depend on the orientation of the source relative to the detector.

Figure 7.8 shows the sensitivity of various gravitational wave observatories. The Laser Interferometer Space Antenna (LISA)[32] is a planned interferometer consisting of three spacecraft whose positions mark the vertices of an equilateral triangle 5 km on a side in an orbit around the Sun. As evident from Figure 7.8, LISA will be able to detect $\sim 10^4$ to $10^7\ M_\odot$ binaries out to arbitrary redshifts during the cosmic dawn. The next-generation ground-based interferometer, Advanced-LIGO,[33] will be sensitive to binaries involving black hole remnants of massive Population III stars (with $\sim 10^2$ to $10^3\ M_\odot$) out to $z \gg 1$.

An electromagnetic counterpart would greatly reduce the positional error and also determine the redshift of the gravitational wave source, enabling its use as a "standard siren" for precision cosmological distance measurements. Electromagnetic radiation could naturally be produced by a circumbinary accretion disk prior to coalescence, through a variety of possible effects: viscous dissipation of gravitational wave energy in the disk might result in a weak electromagnetic transient shortly after the merger, reequilibration of the inner edge of the disk could create an X-ray brightening on a timescale of 10–10^3 yr, and shocks produced by the remnant's recoil might generate electromagnetic reverberations that might take $\sim 10^4$ years to dissipate as enhanced IR luminosity.[34] It is not obvious whether these signals could be distinguished from the much more abundant sources of temporal variability in quasars with a single black hole. Moreover, the luminosity of any circumbinary disk is expected to be significantly reduced by the cavity associated with the decoupling of the black hole binary from the inner edge of the disk in the final stage of inspiral. The disk is not expected to refill the cavity and return to its full luminosity for at least

Figure 7.8 (Continued.) amplitude of LISA $h_{\rm rms}$ from the detector only (dashed) and from the detector combined with the anticipated foreground confusion (dash-dotted), along with the characteristic amplitudes $h_{\rm char}$ of three binary masses M (solid). The locations on each $h_{\rm char}$ curve correspond to the peak amplitude (circle), 1 hour before the peak (filled circle), 1 day before the peak (circle with inscribed cross), and 1 month before the peak (circle with inscribed square) in the observer frame, as well as times of $25r_{\rm Sch}/c$ (square) and $500r_{\rm Sch}/c$ (diamond) before the peak in the source frame. *Middle panel:* Contour plot of the signal-to-noise ratio (SNR) with binary mass and redshift dependence for LISA. *Bottom panel:* SNR contour plot with mass and redshift dependence for Advanced-LIGO. Reprinted Figures 16, 17, and 19 with permission from Baker, J., et al., *Phys. Rev.* **D75**, 4024 (2007). Copyright 2007 by the American Physical Society.

a decade after coalescence. On longer timescales, the portion of the accretion disk that remains bound to the recoiled black hole remnant is expected to be detectable as a kinematically and eventually a spatially offset quasar, although its lifetime is limited by the supply of gas that can remain gravitationally bound to it.[35] Interestingly, the recoil of the remnant black hole allows it to frequently disrupt stars along its path and produce a prompt electromagnetic signal that does not depend on the prior existence of a gaseous disk in the vicinity of the merging binary.[36] Tidal disruption of stars can be observed out to cosmological distances in cases where a relativistic jet is produced.[37]

The expected event rate of massive binary mergers can be calculated based on the halo merger rate predicted by the excursion set formalism (§3.4.1) under various assumptions about the relation between the black hole and halo masses. For reasonable assumptions, LISA is expected to detect many cosmological events per year. The actual detection of these signals would open a new window into the Universe and enable us to trace the hierarchical assembly of black holes in galaxies throughout cosmic history. Moreover, because gravitational waves pass freely through all forms of matter, gravitational wave observatories might discover new populations of black hole binaries that are electromagnetically faint because of their modest mass relative to bright quasars, or because they are enshrouded by gas and dust.