AGN Physics and Models

https://doi.org/10.1017/S0074180900175096 Published online by Cambridge University Press

MODELS FOR VARIABILITY IN AGNS

MARTIN J. REES
Institute of Astronomy
Madingley Road
Cambridge, CB3 0HA, U.K.

1. Introduction

In this talk I shall address three different processes relevant to continuum variability in AGNs. The first two refer to the physical conditions in the regions responsible for the non-thermal emission, and the implications of high brightness temperatures. The third is the distinctive type of flare that results when a star is tidally disrupted by a massive black hole; this process, which merits much further study, it likely to be specially important as a diagnostic of physical conditions in low-luminosity nearby nuclei.

2. How high can jet Lorentz factors be?

It has for been evident for several years that radio jets, at least on the scales probed by VLBI, display bulk flows with Lorentz factors Γ of up to 10. The evidence comes from the brightness temperature limit on any incoherent synchrotron source which avoids a synchrotron self-Compton catastrophe, and directly from VLBI observations of superluminal motions. Gamma-ray data from the EGRET experiment on the Compton Gamma-Ray Observatory offer independent indications of highly relativistic outflow, as has been discussed by several speakers at this meeting. Some sources have been detected as strong gamma-ray sources which vary on timescales as short as days. If the sources were actually as small as a few light days, then no gamma rays would escape (irrespective of the emission mechanism), because the photon density would be so high that pair production would occur. This process can be avoided by highly relativistic outflow, which allows a larger source size, and raises the photon energy threshold for pair productions in the observer's frame. If the gamma rays are relativistically beamed rather than being isotropic, the implied jet luminosity need be no more than 10^{46} ergs/s; on the other hand, isotropic emission would imply uncomfortably high powers of several times 10^{48} ergs/s.

Although the evidence requires $\Gamma \simeq 10$ (and a jet with power Lj up to a few times 10^{46} ergs/s) it is interesting to ask whether Γ could be still higher? The data do not obviously exclude this, though there are well-known theoretical constraints: in particular, Compton drag near the base of the jet is important, especially if the energy is mainly carried by electron-positron pairs (Rees 1984, Phinney 1985). The issue is now more open, however, because of the realisation that magnetic fields

T. J.-L. Courvoisier and A. Blecha: Multi-Wavelength Continuum Emission of AGN, 239–248.

probably play a dominant role, especially near the base of the jets. The wound-up field would collimate the jet. Moreover Li, Begelman and Chiueh (1992) have shown that, in Poynting-dominated jets, the bulk flow is gradually accelerated despite Compton drag. The drag enhances the energy stored in the field, and this energy can be given back to the jet (as in a decompressing spring). There is therefore no reason for feeling too inhibited about envisaging higher Γ. After all, we are quite comfortable with the idea that the relativistic wind from the Crab pulsar may have $\Gamma \simeq 10^4$.

Neither superluminal motions nor gamma-rays yet force us to Lorentz factors much above 10. But there is now another phenomenon that may do – intraday radio variability (IDV). Intensity changes within a single day, with more than 20 per cent amplitude, have been detected in about a quarter of all compact extragalactic sources. (Quirrenbach et al 1989, Witzel 1992). The claimed correlation with variations in the optical band would preclude interpretations in terms of interstellar scintillations, and imply that they were intrinsic to the source. If the source dimensions were limited to ct, these rapid variations would indicate vastly higher brightness temperatures than are compatible with a synchrotron source. In 0917+624 (Qian et al 1991) the "formal" brightness temperature is 2.10^{18} K. To bring this down below the limit set by the Compton catastrophe would require Γ to be at least 100; the relevant emitting region would then be at a distance of $\sim 10^4$ light days, i.e. several parsecs, from the base of the jet. Is this possible?

The general consequences of postulating higher Γ can be readily envisioned simply by considering how various key timescales depend on Γ and on Lj. Let us consider a jet of fixed luminosity and cone angle, and focus attention on a particular point along its length. Relativistic electrons will cool due to inverse Compton scattering of ambient radiation – radiation from the nucleus itself, and from the gas that emits the broad emission lines. The radiation energy density is $\sim L/r^2c$. As viewed in the frame of the jet the energy density is enhanced by a factor $f\Gamma^2$.

The factor f, equal to 1 for isotropic radiation, depends on the angular distribution of the radiation. Scattering of ambient radiation becomes increasingly important for high Γ. The maximum magnetic field is that for which Poynting flux carries all the jet luminosity Lj. When Γ is larger, straightforward Lorentz transformation tells us that the maximum magnetic field in the comoving frame becomes weaker for a given $L_j(\propto \Gamma^{-1})$. Synchrotron lifetimes then become longer, and synchrotron emission consequently becomes very much less competitive with Compton scattering, for which the characteristic lifetime *decreases* for higher Γ. Moreover synchrotron losses also become slow compared to adiabatic cooling: the timescale for adiabatic expansion at a fixed location along the jet, measured in the comoving frame, goes as Γ^{-1} because of the ordinary time dilation. (A long synchrotron lifetime also guarantees that synchrotron self-Compton emission would be unimportant if a substantial part of the jet energy were transported as Poynting flux)

To produce an apparent synchrotron luminosity $\sim 10^{44}$ erg/s might require only 10^{-4} efficiency if $Lj \simeq 10^{46}$ erg/s and the beaming solid angle were (say) 10^{-2}. But for sufficiently high Γ, the synchrotron efficiency, given by $\min[t_{exp}, t_{Comp}]/t_{sync}$, would fall below even such a low value as this. There is therefore an upper limit to Γ, set by the requirement that the synchrotron efficiency is not too low, and that the high-energy output should not exceed the variable radio luminosity by an unacceptable factor. (These two contributions would be beamed and doppler-boosted in an essentially similar way). The presently-claimed IDV data turn out to be marginally consistent with a synchrotron hypothesis, with Γ in the range 100-200. This reduces the motivation for invoking coherent processes (eg Benford 1992). However the speed of these variations is right at the limit: any radio variations that were substantially more rapid could not be accommodated without invoking radiation processes that permit higher brightness temperatures.

Note that the above arguments set an upper limit to Γ specifically for a jet (or a part of a jet) that emits synchrotron radiation above some threshold efficiency. They do not rule out far higher Γ in a jet that is simply a conduit for energy, or which is detectable only via inverse Compton scattering of ambient radiation. It is indeed quite possible that Poynting flux can be carried by very high-Γ jets. Moreover it is possible that in some jets (eg that of M 87) the material near the axis has $\Gamma \gg 100$, the observed synchrotron radiation coming from a violently-sheared boundary sheath that moves more slowly (though perhaps still relativistically).

3. Induced (or stimulated) Compton scattering

Whatever radiation mechanism the IDV sources turn out to involve, they offer a further motive for considering a process which is important in high-surface-brightness sources – induced compton scattering. This process, originally known as the "Dirac-Kapitza effect", has been known for a long time, and causes apparent changes in the spectrum and structure of a source. Its effects are inherently non-linear, and therefore tricky to calculate, but it may well be significant in compact radio sources. Induced scattering dominates spontaneous scattering whenever the occupation number n exceeds unity. At first sight, one would have thought that it should consequently be colossally important. However, if there are two beams with respective occupation numbers n_1 and n_2, induced scatterings from 1 to 2 ($n_1 \times n_2$) are compensated by those in the reverse direction ($n_2 \times n_1$). The cancellation would be exact were it not for the effects of electron recoil, which amounts to a fraction $h\nu/m_e c^2$ of the photon momentum. A necessary condition for induced scattering to be significant is therefore that the photon occupation numbers exceed $m_e c^2/h\nu$; equivalently, a brightness temperatures above $m_e c^2/k$ is required.

The brightness temperatures of many compact radio sources are directly observed by VLBI to be as high as 10^{12}K (this is, of course, well known to be close to the limit allowed by a self-absorbed incoherent synchrotron source). Induced scattering effects could therefore be significant in these source components even if

the Thomson depth in or around them were only $\sim (m_e c^2/kT_b) \simeq 10^{-2}$.

The consequences of induced scattering are frequency-dependent, and depend on the radiation intensity and spectrum in non-linear ways. The simplest effect, first discussed by Sunyaev (1970) and Levich (1972), is a "Bose-condensation" of radiation towards lower frequencies. If induced scattering is too important, the radiation is, in effect, completely absorbed; but for induced optical depths of order unity the low-frequency brightness temperature can actually be enhanced above the self-absorption limit (though generally only by a modest factor), and steep gradients can be created in the spectrum. An ultra-compact high brightness core surrounded by tenuous thermal plasma may consequently produce a spectrum with a second component, due to induced scattering, peaking at a lower frequency than the intrinsic spectrum from the core.

When spatial structure can be resolved, the apparent brightness of each part of the source can be modified by induced scattering; moreover these modulations change in a complicated way when the source varies. If such effects were important, they would confuse the interpretation of variability and superluminal motions (Wilson 1982, Coppi, Blandford and Rees 1993). The firmest indications of induced scattering would be strong spectral variations in source structure and large frequency-dependent linear polarization.

There is no compelling evidence that induced Compton effects have been observed in any compact radio source. This already sets non-trivial upper limits on the density of thermal plasma within the central few parsecs of active galaxies. Future observations with VLBI arrays will have greater dynamic range; detection of such effects would offer a new probe for physical conditions in AGNs. Blandford (1993) has pointed out that stimulated Raman scattering off collective plasma waves, which has similar observational characteristics, may be even more important.

Induced scattering could be catastrophic in coherent sources. For instance, suppose that the region at the base of the jets, where the field strengths may be $\sim 10^4$ G and the gyrofrequency at centimetre wavelengths, emitted coherent cyclotron radiation. (This is not an absurd hypothesis, and would certainly remove the problem of explaining intraday variability.). But the brightness temperatures would then be $\gtrsim 10^{20}$ K, and induced scattering would quench the spectrum unless τ_{es} were below 10^{-8}. Since τ_{es} is characteristically of order unity, this would seem to require a rather special geometry.

4. Flares from tidally-disrupted stars, especially in low-luminosity nuclei

Tidal disruption of stars was investigated in the 1970s as a possible fuelling mechanism for AGNs. However, it became clear (eg Frank 1978) that this process was unlikely to be significant in high-luminosity objects – to get a high enough disruption rate, the stellar concentration would have to be so extreme that stellar

collisions would supply even more fuel. Tidal disruption is a relatively more significant process in low-luminosity AGNs. Indeed it is of greatest interest in quiescent galactic nuclei. If one accepts that massive black holes lurk in the centres of many (maybe even most) galaxies, then gas accretion must, in these systems, be suppressed, and it is not difficult to think of how this might come about. On the other hand, the occasional deflection of a star onto a near-radial orbit, bringing it within the Roche radius of the massive hole, seems unavoidable. It is therefore of interest to explore the consequences of such events: these may be manifested as occasional flares in quiescent galaxies, thereby allowing us to test the hypothesis that massive black holes indeed lurk in such places.

Evidence is strengthening, from studies of the central stellar velocities, for dark central mass concentrations in the nuclei of several nearby galaxies (see Kormendy 1993 for a recent review). For example, the centre of M 31, even less active than that of our own Galaxy, may contain a black hole of $\sim 3.10^7$ $M_\odot$. The quiescence implies that accretion of gas is proceeding very slowly, and/or is very inefficient radiatively. However, there would be occasional tidal disruption of stars that diffused onto nearly radial orbits. The rate of such events depends on the parameters of the central star cluster. It is therefore somewhat uncertain, but is a "clean" problem involving stellar dynamics rather than gas dynamics. Simple estimates based on models for the stellar distribution in the innermost few parsecs suggest a rate of around 10^{-4} per year. (Now that the HST is providing more detailed information about the central stellar distribution, it would be worthwhile making better estimates)

If this capture rate is translated into a time-averaged accretion rate, it could potentially yield a luminosity higher than is seen from the nucleus of M31. This is not a contradiction, provided that the debris is swallowed or ejected in a flare which lasts only a small fraction of the interval between successive events. What would be the "light curve" and the spectrum of such a flare, and the likely energy output in kinetic form? These depend on the complicated question of how the material, falling in an unsteady, non-axisymmetric fashion, gets accreted, and on the effects of radiation pressure during phases when the dissipation rate is "super-Eddington".

4.1. THE FATE OF THE DEBRIS

Earlier investigations of this phenomenon (eg Rees, 1988, Evans and Kochanek 1989, Canizzo et al 1990) have led to the following inferences:

(i) The actual disruption process is inconspicuous. Tidal forces induce violent shocks, which more than double the star's internal temperature so that its self-gravity cannot prevent it from flying apart. However, the radiation thereby generated is trapped within the star and would be attenuated by adiabatic cooling before being able to escape. (The situation is similar to a supernova explosion, which would not appear bright if there were only a "prompt" energy input at the time of the explosion itself.),

(ii) The debris in the most tightly-bound orbit reaches apocentre and falls back after a few months. This material acquired, during the tidal disruption, a velocity deficit of order $v_* \simeq (Gm_*/r_*)^{\frac{1}{2}}$ relative to the star's centre of mass, and consequently an energy deficit larger than the star's binding energy by the ratio of the orbital speed at pericentre to v_*.

(iii) At much later times the infall rate declines in proportion to $t^{-5/3}$ as more loosely bound material returns to the hole. About half the debris moves on orbits bound to the hole; the rest is on hyperbolic orbits, which escape with speeds up to $\sim 10^4$ km/s.

What is much less obvious is what happens to the debris after it completes its first orbit? Does it quickly circularise? And how quickly and efficiently is radiation emitted when it accretes onto the hole?.

A tidally disrupted star, as it moves away from the hole, develops into an elongated banana-shaped structure, the part on the most tightly-bound orbit (the first to return to the hole) being at one end (Laguna et al 1993, Kochanek, 1993, Lee and Monaghan 1993). If the debris were moving in a Newtonian "1/r" potential, then in principle it could wind up into an elliptical spiral, where successive turns would eventually merge to make an elliptical accretion disc. Such a disc could survive for many orbits if its viscosity (or α-parameter) were low enough (cf Gurzadyan and Ozernoi 1980, Syer and Clarke 1992). It would not then yield a spectacular flare; on the other hand, the decay timescale could be so slow that one would expect to see evidence of a residual disc in most galactic nuclei containing black holes

But the possibility that the debris could neatly wind up into an elliptical disc is precluded by relativistic precession. Consider first the case of a Schwarzschild hole. Orbits then remain in a plane, but each pericentre passage precesses the orbital axis through an angle $\theta_p \simeq 3\,(r_g/r_T)$, where $r_g = GM_h/c^2$ is the hole's gravitational radius. This means that there can be high-speed collisions between bits of debris that have completed different numbers of pericentre passages. This "self-intersection" process begins soon after the most tightly bound debris begins its second orbit: this material can run into inward-falling material returning towards the hole for the first time. Kochanek (1993) and Lee and Monaghan (1993) have tracked the debris to the stage when this encounter starts.

The relative velocity v_{rel} between the outgoing and incoming material is high – enormously supersonic with respect to its internal sound speed. All the debris has essentially the same specific angular momentum around the hole. If the encounter were dissipative, then the coalesced material would have an orbit that was much tighter (and more nearly circular). In fact, even though the post-shock densities are high and the radiative cooling timescales short, the shock is actually nearly elastic. This is because, owing to the high density, the electron-scattering optical depth is large compared with c/v_{rel}; before it can escape, the resultant radiation therefore gets adiabatically cooled, converting the dissipated energy into expansion of the shocked material, which sprays out in a range of directions with speed $\sim v_{rel}$. The shocked material would have a broad spread in angular momentum relative

to the hole: it would not be confined narrowly to the original orbital plane, and a substantial fraction would be counter-rotating with respect to the original orbit. This material would be reshocked before it had completed more than one further orbit, and would circularise within a time shorter than the period of the original elliptical orbit.

There is a second self-intersection point near pericentre, when material returning for the first time meets other material on a more tightly bound orbit returning for the second time. The angle between the two streams is then $\sim \theta_p$. This is a relatively small angle; but since the velocities are largest at pericentre the energy dissipated can still be significant.

A further complication arises if (more realistically) the hole is described by a Kerr rather than Schwarzschild metric. The orbital angular momentum of the captured star will generally be mis-aligned with the hole's spin. Owing to Lense-Thirring precession, the debris will not then remain in a plane. This can in some cases prevent self-intersections; on the other hand, the relative velocity of the collisions can be increased. Overall, this complication provides a further parameter which would render these tidal-capture events highly non-standardised in their outcomes.

After being rendered axisymmetric, the bound debris could, without redistribution of its angular momentum, settle into a torus whose density maximum would be at a radius $\sim 2r_T$, corresponding to an orbital period of only a few hours. If there were no viscosity to redistribute angular momentum, it would cool into a thin ring. However, it is more likely that the "alpha-viscosity" is high enough to maintain it as a thick torus, and allow accretion into the hole – all that this requires is $\alpha \gtrsim 10^{-3}$. [There is, moreover, a special reason why, in this situation, a very low viscosity is unlikely: the Bardeen-Petterson effect would twist the disc around a Kerr hole, and there would be an extra transfer of angular momentum associated with this non-axisymmetric effect.]

4.2. THE "LIGHT CURVE"

These processes are too complex and poorly modelled to permit any confident inferences. However, I believe one can with fair confidence argue that the luminosity is determined by the rate at which debris is "raining down" onto the hole after completing one orbit – in other words, there is not a substantial timelag between the "first return" of the debris and its accretion or ejection. When the infall rate exceeds what is needed to produce the Eddington luminosity for high efficiency, L will be of order L_{Ed}. However, it is harder to assess whether the surplus mass is then expelled by radiation pressure, or swallowed with low efficiency. The location of the effective photosphere, and hence also the spectrum and bolometric correction, depend on the answer to this question, and therefore remain uncertain. The radiation would be predominantly thermal, with a temperature of order 10^5K; however the energy dissipated by the shocks that occur during the circularisation process would provide an extension into the X-ray band.

As illustrative examples, suppose a solar-type star passes just within the tidal radius r and is tidally disrupted by a hole of mass M_h, for the cases $M_h = 10^6 M$ and $M = 10^8$ $M_\odot$. The orbital period for the most tightly bound debris depends on $M_h^{\frac{1}{2}}$, and is respectively 2 months and 2 years in these two cases. In the first case, the infall rate is "super-Eddington" for about a year. We would therefore expect a luminosity to stay at a "plateau" level (though perhaps with a changing spectrum) throughout that period, and then to fade as $t^{-\frac{5}{3}}$. In the higher-mass (10^8 $M_\odot$) case, the accretion rate never gets high enough to generate L_{Ed}, so the luminosity would smoothly rise and fall with a timescale of 1-2 years, without there being necessarily any radiation-driven outflow.

The events would not have standardised properties. They would depend on the impact parameter: in particular, for passages much closer than r_T the timescale is shorter. They also depend on the type of star captured, and the spin of the hole. Observed effects may also depend on orientation, the amount of reprocessing, and the effects of directionality, etc.

Supernova-type searches with $\gtrsim 10^4$ galaxy-years of exposure should either detect instances of this phenomenon, or else place limits on its nature. Such evidence would also help to test whether every galaxy has been through an AGN phase, leaving a 'fossil' black hole at its centre (cf Haehnelt 1993). The possibility of such a bonus should be an added motivation for groups engaged in such searches.

4.3. STARS IN RELATIVISTIC ORBITS?

A separate question, related somewhat to the issue of tidal disruption, is whether a star can be captured into a tightly bound orbit around a massive black hole without being destroyed. Orbits just outside the tidal radius would be close enough to the hole to manifest interesting relativistic effects, such as have been computed in detail by Karas and Vokrouhlicky (1993a,b).

It is readily shown (eg Rees 1988) that a star cannot reach such an orbit by the kind of "tidal capture" process that can create close binary star systems. This is because the binding energy of the final orbit is far higher when the companion is a supermassive hole than when it is also of stellar mass, and this energy cannot be dissipated by the star without destroying it: a star whose orbit brings it within (say) $3r_T$ of a massive black hole may not be destroyed on first passage, but if it is then on a bound elliptical orbit, it will surely get puffed up and disrupted before the orbit has circularised. However, Syer, Clarke and Rees (1991) pointed out a different mechanism. A star's orbit can be "ground down" by successive impacts on a disc (or any other resisting medium): the orbital energy lost does not, in this case, have to be radiated by the star. Other constraints on the survival of stars in the hostile environment around massive black holes – tidal dissipation when the orbit is eccentric, irradiation by ambient radiation, etc – are explored by Podsiadlowski and Rees (1993) (See also King and Done (1993)). Such stars would not be directly observable, though they might cause quasiperiodic modulation

of the AGN emission. Such a phenomenon could offer important information on strong-field gravity, as well as on the geometry of the continuum-emitting region in AGNs; it is therefore well worth searching for.

5. Concluding comments

The three topics I have outlined here all need (and offer timely opportunities for) theoretical clarification and development. The collimation and generation of jets is a challenging problem in relativistic MHD; the variability stems probably from internal shocks quite far out in the jet, which are triggered either by variability at the base, or by interaction with "obstacles". Radiative transfer when induced scattering must be taken into account (section 3) is sufficiently complicated that even the simplest "test problems" await a proper treatment. The phenomenon of tidal disruption, discussed in section 4, offers a specially daunting computational challenge – it involves relativistic, non-axisymmetric unsteady gas dynamics and radiative transfer, with an exceptionally large dynamic range.

I am grateful to Mitch Begelman, Roger Blandford, Paolo Coppi and Marek Sikora for collaboration on topics briefly summarised in this talk.

6. References

Begelman, M.C., Sikora, M. and Rees, M.J. 1993 MNRAS submitted
Benford, G. 1992, Astrophys. J. (Lett.), 391, L59 Blandford, R.D. 1993 in preparation.
Canizzo, J.K., Lee, H.M. and Goodman, J 1990 Astrophys. J. 351, 38
Coppi, P., Blandford, R.D. and Rees, M.J. 1993 MNRAS 262, 603.
Evans, C.R. and Kochanek C.S. 1989 Astrophys. J. (Lett) 346, L13.
Frank, J. 1978, MNRAS 184, 87
Gurzadyan, V.G. and Ozernoi, L.M. 1980 Astron. Astrophys. 86, 315.
Haehnelt, M. 1993, these proceedings.
Karas, V., and Vokrouhlicky, D., 1993a MNRAS in press
Karas, V., and Vokrouhlicky, D., 1993b Astrophys. J in press
Khokhlov, A., Novikov, I.D. and Pethick, C.J. 1993 MNRAS in press
King, A.R. and Done, C. 1993 MNRAS 264, 388
Kochanek, C.S., 1993 Astrophys. J. in press
Kormendy, J. 1993 in "Testing the AGN Paradigm" ed Holt, S (AIP Conference Proceedings)
Laguna, P., Miller, W.A., Zurek, W.H., and Davies, M.B. 1993 in "Testing the AGN Paradigm" ed S.Holt (AIP Conference Proceedings)
Lee, H.M. and Monaghan, J.J. 1993 in "Nuclei of Nearby Galaxies" ed Genzel, R and Harris, J. (Kluwer, in press).

Levich, E.V. 1972 Sov. Phys. JETP 34, 59.
Li, Z-Y, Begelman, M.C. and Chiueh, T., 1992. Astrophys. J. 384, 567.
Phinney, E.S. 1985 in "Astrophysics of Active Galaxies and Quasi-Stellar Objects" ed J. Miller p453 (University Science Books, California)
Podsiadlowski, P. and Rees, M.J. 1993 in preparation.
Quirrenbach, A. et al 1989 Astr. Astrophys. 226, L1
Qian, S.J. et al 1991 Astr. Astrophys. 241, 15
Rees, M.J. 1984 in "Very Long Baseline Interferometry" eds R. Fanti et al p.207 (Reidel, Holland).
Rees, M.J. 1988 Nature 333, 523
Sunyaev, R.A. 1970 Astrophys. Lett 7, 19.
Syer, D and Clarke, C.J. 1992 MNRAS 255, 92 (errata in 260, 463)
Syer, D., Clarke, C.J. and Rees, M.J. 1991 MNRAS 250, 505.
Wilson, D.B. 1982 MNRAS 200, 881
Witzel, A. 1992 in "Physics of Active Galactic Nuclei", eds Wagner, S., and Duschl, W. (Springer-Verlag, Berlin)